REV-00

SELF-LEARNING MATERIAL

UNIVERSITY OF SCIENCE & TECHNOLOGY MEGHALAYA

Unveiling Excellence

MASTER OF BUSINESS ADMINISTRATION (HRM)

MBAH 202 : MODULE ON QUANTITATIVE TECHNIQUES

w.e.f Academic Session: 2023-24

CENTRE FOR DISTANCE AND ONLINE EDUCATION

UNIVERSITY OF SCIENCE & TECHNOLOGY MEGHALAYA

nirf India Ranking-2023 (151-200)

Accredited 'A' Grade by NAAC

Techno City, 9th Mile, Baridua, Ri-Bhoi, Meghalaya, 793101

SELF-LEARNING MATERIAL

Master of Business Administration (HRM)

MBAH 202

Module On Quantitative Techniques

Academic Session: 2023-24

Unveiling Excellence

Centre for Distance and Online Education

UNIVERSITY OF SCIENCE & TECHNOLOGY MEGHALAYA

Accredited 'A' Grade by NAAC

Self Learning Material
Center for Distance and Online Education
University of Science and Technology Meghalaya

First Edition
Print Jan 2024

Edited by: Mr. Jayanta Sarma Kakoty

This book is a distance education module comprising of collection of learning material for students of Center for Distance and Online Education, University of Science and Technology Meghalaya, 9th Mile G S Rd, Ri Bhoi, Meghalaya 793101.

Printed and Published on behalf of Center for Distance and Online Education, University of Science and Technology Meghalaya by Publication Cell, University of Science and Technology Meghalaya – 793101

MBAD 202
QUANTITATIVE TECHNIQUES AND STATISTICS IN BUSINESS

CONTENTS

UNIT 1: INTRODUCTION TO STATISTICS

Structure

The word 'Statistics' has been derived from the Latin word 'Status' or the Italian word 'Statista' or the German word 'Statistik'. The meaning of these words are same, which is 'Political State'. In ancient times, 'Statistics' was related to achieve political purposes of rulers – to assess the manpower and to introduce new taxes or levis. The Governments used to collect information regarding the population and property of wealth of the state. The science of Statistics developed gradually and its field of application widened day by day. With the introduction of theory of the study of Agriculture, Biology, Electronics, Medicines, Political Science, Economics, Psychology, Sociology, Business and Commerce.

1.1 Meaning and definition of Statistics

Meaning of Statistics:

The word 'Statistics' is used in two different senses – Plural and Singular.

In plural form 'Statistics' means collection of numerical statements of facts or quantitative data pertaining to a phenomenon such as number of persons suffering from malaria in different areas in Meghalaya or the number of unemployed girls in different states in India and so on.

In singular form 'Statistics' means methods adopted for collection, organization, presentation, analysis and interpretation of quantitative data.

Definitions of Statistics:

Definition of Statistics in Plural form or Statistics as Statistical data:
Different authors have given different definitions of Statistics. A comprehensive definition was given by Horace Secrist, which is as follows:
"Statistics are the aggregate of facts affected to a marked extent by multiplicity of causes, numerically expressed, enumerated or estimated according to reasonable standards of accuracy collected in a systematic manner for a pre-determined purpose and placed in relation to each other."

Definition of Statistics in Singular form or Statistics as Statistical methods:
The definitions of Statistics used in singular form i.e., Statistics as statistical methods were given by some authors which are given below:

According to **Prof. Bowley**, "Statistics may be called the Science of counting." This definition covers only one aspect i.e., counting, but the other aspects such as organization, presentation, analysis and interpretation have been ignored.

According to **Prof. Boddington**, "Statistics is the Science of estimates and probabilities." This definition is partially accepted as it does not cover all the aspects of Statistics.

According to **Berenson and Levin**, "Statistics, as a Science can be viewed as application of scientific method in the analysis of numerical data for the purpose of making rational decisions." This definition covers only analysis of data, it ignores the other aspects.

Croxton and **Cowden** has defined "Statistics' as follows:
"Statistics may be defined as the collection, presentation, analysis and interpretation of numerical data." This definition covers almost all the aspects of 'Statistics', but ignores one aspect, that is, "organization of numerical data."

Hence, based upon the above discussions, it is concluded that " **The word 'Statistics' is**

defined as collection, organization, presentation, analysis and interpretation of numerical data.”

1.2 Characteristics of Statistical data:

From the definition given by **Horace Secrist**, it is clear that ‘Statistics’ should possess the following characteristics:

(i) **Statistics are aggregate of facts:** This means that a single and isolated fact or figure cannot be Statistics.

(ii) **Statistics must be numerically expressed:** It means that quantitative data like heights, weights, prices, quantities, number of persons etc. will constitute Statistics.

(iii) **Statistics should be capable of being related to each other:** It means that Statistics should be capable of comparison and connected with each other of same category.

(iv) **Statistics should be collected for a predetermined purpose**: It means that before starting with collection of data, purpose or objective of collection should be designed and work should be done accordingly.

(v) **Statistics are affected to a marked extent by multiplicity of causes**: It means that the data should be liable to considerable changes in their values at different times, places or situations and the changes in the values of the data should be the result of interaction of a number of factors.

(vi) **Statistics must be enumerated, estimated or approximated according to reasonable standard of accuracy**: It means that if the group of numerical data are huge amount, without counting each and every time, a sample is drawn from this group and the results obtained from the sample are estimated or approximated to the whole group.

1.3 Applications of inferential statistics in managerial decision making

The ‘Statistics’ can be divided into the following two categories.

(i) Descriptive Statistics and

(ii) Inferential Statistics

By **Descriptive Statistics** we mean those statistical methods which are used in the collection, organization, presentation, analysis and interpretation of data. The most commonly used methods of statistical analysis include measures of central tendency, measures of dispersion (variation), moments, skewness, kurtosis, correlation and regression etc. **Descriptive Statistics** also involve various statistical tools such as bar charts, histograms, frequency curve, ogives etc. On the other hand, **Inferential Statistics** include those statistical methods which facilitate the estimation of various characteristics of a population or making decisions concerning a population on the basis of sample results. **Population** and **Sample** are relative terms.

An aggregate of items under study is called **population** or **universe**. A finite subset of a population or a universe is called a **sample**. The number of individuals in a sample is called the **sample size**.

Inferential Statistics mainly work on two factors: **statistical estimation** and **hypothesis testing**.

In **statistical estimation**, we estimate the value of the population parameter from sample statistics. In **hypothesis testing**, we test some hypothesis about parent population from which the sample is drawn.

In many situations, we collect data from a large group of elements such as companies, households, products, customers, voters etc. But, it is not possible to draw conclusion based upon the information obtained from the large population. Therefore, we study a group of individuals, called sample and draw a conclusion about a population on the basis of information contained in a sample drawn from that population.

This helps the decision maker to draw conclusions about the characteristic of a large population under study.

With the growth in the size of business firms, it is quite impossible for the decision makers to maintain personal contact with the thousand of customers. A manager has to plan, organize, supervise and control the whole operations of any business activity. But due to little contact with the customers he / she faces a great degree of uncertainty regarding the future policies.

Statistical reports provide a summary of business activities which improve the capability of making more effective decisions regarding future activities. Below are some major areas where a businessman can make use of statistics for making good decisions.

I. **Marketing**: Before a product is launched, the manager of a firm, along with his team makes a pilot survey to analyse data on various factors such as purchasing power of customers, habits of customers, number of competitors, pricing of other products etc. by using the various statistical techniques. Such studies help the manager in knowing the possible market potential for the product. Therefore, by using the concerned statistical techniques, the manager can make advertising strategies, routing of salesman, establish sales territories and hence improve the sales of the product.

II. **Production**: By using statistical methods, a manager can carry research and development programmes for improving the quality of existing products and set quality control standards for coming products. This will help the manager in taking the decisions about the quantity and the time of either self-manufacturing or buying from outside.

III. **Finance**: By studying the correlation analysis of profits and dividends, a firm manager can predict and decide the probable dividends for coming years. Similarly, the analysis of data on assets and liabilities, income and expenditure, sales and purchase etc., can help the manager to ascertain the financial results of various operations.

IV. **Manpower planning**: For smooth functioning of big organizations, a proper manpower planning is essential. A manager can make use of statistical studies of wages, incentive plans, cost of living, labour turnover rates, employment trends, accidents
Rates, training and development programmes, employer-employee relationships etc. This will help the manager in formulation future policies and plans for the overall success of an organization.

Hence, statistics and statistical methods can provide the businessperson with one of his / her most valuable tools for decision making.

1.4 Limitations of Statistics

Statistics has some limitations, which restricts its scope and importance. The following are the important limitations of Statistics.

(i) **Statistics does not deal with individual items**: Statistics deals with groups or aggregates only and the study of an individual fact lies outside the scope of Statistics.

For example, the weight of a student is 50 kg, does not constitute Statistics, whereas the average weight of a student in a certain class, will constitute Statistics.

(ii) **Statistics does not deal with qualitative data**: Statistics does not study the data which cannot be measured in quantitative form. For example, average height of students of a class, per capita income of a Country etc. can be studied by the statistical methods. But qualitative aspects such as intelligence, poverty, blindness, deafness, honesty etc. cannot be studied directly.

(iii) **Statistical laws are true only on averages**: Statistical results are not absolutely true. They are not applicable to all individual cases, although they are derived they are derived as the average result of all the individuals forming the groups. For example, the average marks of a group of students does not mean that each student of the group has secured the same average mark.

(iv) **Statistics does not reveal the entire story of a problem:** Since most of the problems are affected by such factors which are incapable of statistical analysis, it is not always possible to examine a problem in all manifestations only by a statistical approach. Many problems have to be examined in the background of a country's culture, philosophy or religion. All these things do not come under the domain of Statistics.

(v) **Statistics is liable to be misused**: Since, Statistics deals with figures and figures are flexible and can be easily distorted, manipulated or moulded by the inexpert and unskilled workers or by the motivated, dishonest and unscrupulous persons, it is very much likely to be misused in most of the cases.

(vi) **Statistical laws are not perfectly accurate**: Statistics deals with the phenomena which are affected by multiplicity of causes and it is not possible to study the effects of each of these factors individually as it is done by experimental methods.

(vii) **Statistical data should be uniform and homogeneous**: One of the important characteristics of statistical data is comparison. Uniform and homogeneous data can be compared. Unequal or uncomparable will lead to wrong and misleading results.

1.5 Classification and Tabulation

Classification of data:

Classification is a process of arranging data into sequences and groups according to their common characteristics or separating them into different but related parts.

Importance / purposes of classification:

(i) Classification condenses the mass of data in such a way that salient features can be readily noticed.

(ii) It facilitates comparisons between attributes of variables.

(iii) It prepares data which can be presented in tabular form.

(iv) It highlights the significant features of the data at a glance.

Types of classification

There are four types of classification of data, which are explained below:

(i) **Geographical Classification**: In this classification, data are classified on the basis of geographical or locational differences such as – cities, districts or villages between various elements of data set.

(ii) **Chronological Classification**: When data are classified on the basis of time, the classification is known as chronological classification. Such classifications are also called time series because data are usually listed in the chronological order starting with the earliest period.

(iii) **Qualitative Classification**: If the data are classified according to the characteristics having no numerical figure i.e., attributes like sex, literacy, religion, cast or education etc., then the classification is known as qualitative classification. This classification is done either single attribute called simple classification or two or more attributes called manifold classification.

(iv) **Quantitative Classification**: In this classification, data are classified according to the characteristics having numerical figures i.e., variable such as height, weight, income, expenditure, productions or sales. This classification is based on either discrete or continuous.

Characteristics of an Ideal Classification

(i) **It should be unambiguous**: It is necessary that various classes should be so defined that there is no room for confusion. There must be only one class for each element of the data set.

(ii) **Classes should be exhaustive and mutually exclusive**: Each element of the data set must belong to one and only one class.

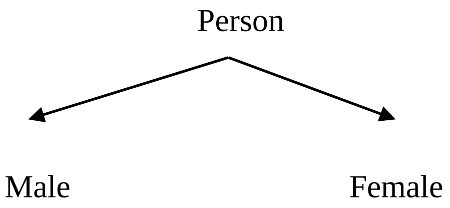
(iii) **It should be stable**: It means that if each time an investigation is conducted, it remains unchanged and hence the results of one investigation may be computed with that of the other.

(iv) **It should be flexible**: It means that suitable adjustment can be made in new situations and circumstances.

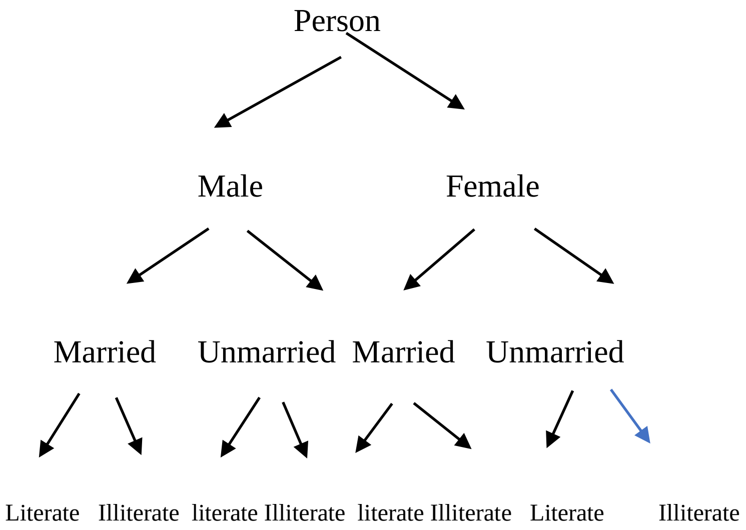
Methods of Classification of statistical data:

There are two methods of classification of statistical data viz., (i) Classification according to attribute and (ii) Classification according to variable.

(i) **Classification according to attributes**: In this classification, data are classified according to the characteristics which cannot be expressed numerically. If the data are classified according to the single attribute, then it is called simple classification. For example, if we classify the a person according to sex, then it is simple classification.



Again, if we classify data according to two or more attributes, then it is called manifold classification. For example, if we classify the data according to sex, then marital status and literacy, then it is manifold classification.



(i) Classification according to variables: In this classification, data are classified according to the characteristics which can be expressed numerically such as heights, weights, marks, incomes, expenditures, productions, or sales etc. Here, the classification is based on either discrete or continuous.

In **discrete classification**, the data are classified according to their magnitudes along with their corresponding frequencies. For example:

Marks	: 5	15	25	35	45
No. of students:	9	20	35	10	3

In **continuous classification**, the data are classified into the number of groups or classes, each of which is called class interval along with their corresponding class frequencies. For example:

Marks	: 0 – 10	10 – 20	20 – 30	30 – 40
No. of students:	5	15	20	3

Tabulation of data

Tabulation may be defined to be orderly and systematic presentation of numerical data in rows and columns designed to clarify the problem under consideration and to facilitate comparison between the figures.

Importance / purposes of Tabulation of data

(i) It simplifies the complex data.
(ii) It requires economic space to present the data
(iii) It facilitate comparison of statistical data
(iv) It helps reference for future needs.

Types of Table

The different types of table, are explained below.

(i) **General and specific purpose table**: It deals with general economic conditions and can be used for various purposes.

(ii) **Original and derived tables**: original tables contain actual and original data. Derived tables contain derived results from the original data.

(iii)Process tables: These tables help processing of data for analysis and interpretation**.**

(iv)**Summary tables**: These tables provide summary results of an enquiry at the end of enquiry.

(v) **Simple and complex tables**: A simple table deals with only the sub-classes of a given phenomenon related with some other variables, whereas a complex table deals with more involved of one or more variables and is related with various sub-sections of another variable.

Different parts of a table

A table consists of the following parts

(i) **Number and title**: The serial number of the table and the subject matter of the table.

(ii) **Stubs**: The heading of rows.

(iii) **Captions**: The headings of the columns

(iv) **Body**: The figures in the table.

(v) **Foot-note**: Source from which the data have been obtained.

Serial Number:

Title:

Column → **Row** ↓	**Caption**	**Caption**	**Caption**	**Total**
Stub	B			
Stub		O		
Stub			D	
Total				Y

Foot-note:

Characteristics / Essentials of a good / ideal Table:

(i) A suitable heading should be given to the table. It should be brief, comprehensive and self-explanatory.

(ii) Headings of columns and rows should be brief and clear.

(iii) The rows and columns should arranged in logical order.

(iv) Figures to be compared, should be placed as near to each other as possible.

(v) The table should be self-explanatory and as simple as possible.

(vi) Explanatory notes and sources from which the data are obtained should be given as foot-notes.

(vii) The units of measurements under each heading or sub-heading must always be indicated.

(viii)As far as possible, the figures should be approximated before tabulation.

1.6 Measures of Central Tendency (Averages):

In many frequency distributions, the tabulated values show small frequency at the beginning and at the end and very high frequency at the middle of the distribution. This indicates that the typical values of the variable lie near the central part of the distribution and other values cluster around these central values. This behavior of the data about the concentration of the values in the central part of the distribution is called central tendency of the data. We shall measure this central tendency with the help of mathematical quantities**. A central value which 'enables us to comprehend in a single effort the significance of the whole' is known as Measure of central Tendency or Statistical Average or simply Average.** In fact, an average of a statistical series is the value of the variable which is representatives of the entire distribution and therefore, gives a measure of central tendency.

Importance/ Usefulness of Measures of Central Tendncy (Averages):

(i) **It is useful to extract and summerise the characteristics of the entire data set in a precise form.** For example, it is difficult to understand individual families need for water during summers. Therefore, knowledge of average quantity of water needed for entire papulation will help the government in planning for water resources.

(ii) **It facilities comparison between two or more data sets.** Such comparison can be made either at a point of time or over a period of time. For example, average sales figures of any month can be compared with the preceding months, or even with the sales figures of competitive firms of same months.

(iii) **It offers a base for computing various measures such as dispersion, skewness, kurtosis that help in many other phases of statistical analysis.**

Requisites or characteristics of an ideal Measure of Central Tendency (Averages):

The following are the characteristics to be satisfied by an ideal measure of central tendency

(i) It should be rigidly defined.

(ii) It should be readily comprehensible and easy to calculate.

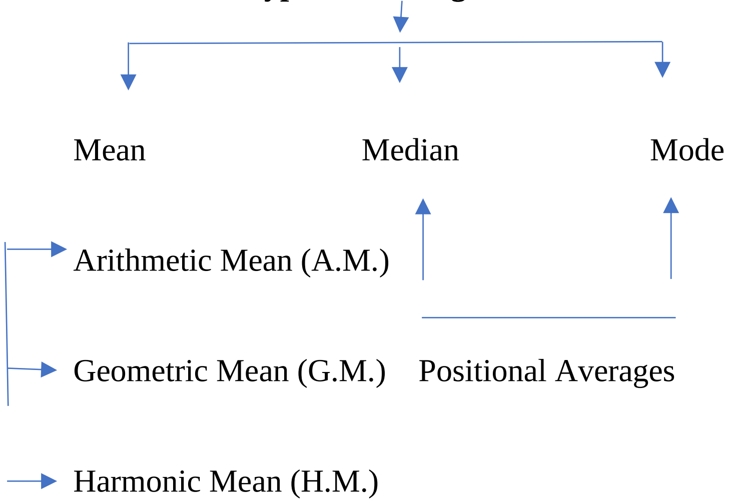
(iii) It should be based on all observations.

(iv) It should be suitable for further mathematical treatment.

(v) It should be affected as little as possible by fluctuations of sampling.

(vi) It should not be affected much by extreme values.

Types of Average / Measures of Central tendency



Arithmetic Mean (A.M.)

In case of simple or ungrouped data, if $x_1, x_2, x_3, \ldots, x_n$ be the values of the variable X (say), then A.M., is defined as A.M. = $\frac{Total\ values}{Number\ of\ values}$ and it is denoted by$\bar{X}$.

Thus, $\bar{X} = \frac{x_1+x_2+x_3+\cdots+x_n}{n}$. Symbolically, it is expressed as $\boldsymbol{\bar{X} = \frac{\sum X}{n}}$

In case of grouped data, if $f_1, f_2, f_3, \ldots, f_n$be the corresponding frequencies of the values $x_1, x_2, x_3, \ldots, x_n$ of the variable X, then the A.M. is given by $\bar{X}$

Thus, $\bar{X} = \frac{x_1+x_2+x_3+\cdots+x_n}{n}$. Symbolically, it is expressed as $\bar{\boldsymbol{X}} = \frac{\sum \boldsymbol{X}}{\boldsymbol{n}}$

In case of grouped data, if $f_1, f_2, f_3, \ldots, f_n$be the corresponding frequencies of the values $x_1, x_2, x_3, \ldots, x_n$ of the variable X, then the A.M. is given by

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

Or, $\bar{\boldsymbol{X}} = \frac{\sum \boldsymbol{fx}}{\sum \boldsymbol{f}}$

Ex.1. (i) Find the A.M. of the following numbers:

5, 8, 10, 15, 24 and 28

(ii) Find the A.M. of the following series:

4, -2, 7, 0 and -1

Sol. (i) A.M. $\bar{\boldsymbol{X}} = \frac{\sum \boldsymbol{X}}{\boldsymbol{n}} = \frac{\mathbf{5+8+10+15+24+28}}{\mathbf{6}} = 15$

(ii) A.M. $\bar{\boldsymbol{X}} = \frac{\sum \boldsymbol{X}}{\boldsymbol{n}} = \frac{\mathbf{4-2+7+0-1}}{\mathbf{5}} = 1.6$

Ex.2. Find A.M.

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

Sol.

Age (x)	No. of students (f)	fx
20	1	20
19	2	38
18	4	72
17	8	136
16	11	176
15	10	150
14	7	98

13	4	52
12	2	24
11	1	11
Total	$\boldsymbol{\sum f = 50}$	$\boldsymbol{\sum fx = 777}$

∴A.M. $\bar{X} = \frac{\Sigma fx}{\Sigma f} = \frac{777}{50} = 15.54\, years$

Ex.3. Determine A.M.

Marks obtained:	**0 - 10**	**10 – 20**	**20 – 30**	**30 – 40**	**40 – 50**
No. of students:	**6**	**5**	**8**	**7**	**4**

Sol.

Marks obtained	**f**	**Mid value x**	**fx**
0 – 10	6	$\frac{0+10}{2} = 5$	30
10 – 20	5	$\frac{10+20}{2} = 15$	75
20 – 30	8	$\frac{20+30}{2} = 25$	200
30 – 40	7	$\frac{30+40}{2} = 35$	245
40 – 50	4	$\frac{40+50}{2} = 45$	180
Total	$\sum f = 30$		$\sum fx = 730$

A.M. $\bar{X} = \frac{\Sigma fx}{\Sigma f} = \frac{730}{30} = 24.33\, marks$

Ex.4. Calculate arithmetic average of wages of labourers an enterprise from the following distribution.

Wages (in Rs):	**15 – 25**	**25 – 35**	**35 – 45**	**45 – 55**	**55 - 65**	**65 – 75**
No. of labourers:	**4**	**11**	**19**	**14**	**5**	**2**

(Do yourselves)

Properties of Arithmetic Mean

(i) Sum of the deviations of the values of the variable from the A.M. is zero

i.e, $\sum(X - \bar{X}) = 0$ (Simple / ungrouped data)

or $\sum f(X - \bar{X}) = 0$ (Grouped data)

(ii) Arithmetic Mean is not independent of the change of origin. i,e, if $\bar{X}$ be the A.M. of the variable X, the arithmetic mean of $(X \pm a)$is $\bar{X} \pm a,\ a \neq 0$

(iii)Arithmetic Mean is not independent of the change of scale. i,e, if $\bar{X}$ be the A.M. of the variable X, the arithmetic mean of $X \times a\ is\ \bar{X} \times a$ and the arithmetic mean of $\frac{X}{a} is \frac{\bar{X}}{a}, a \neq 0$

Combined Mean

Let, $n_1$and n_2 be the number of observations of two sets of data say X_1 and X_2. Let, $\bar{x}_1$ and $\bar{x}_2$be their respective means (A.Ms). Therefore, we have

	$\boldsymbol{X_1}$	$\boldsymbol{X_2}$
Number of observations	n_1	n_2
Mean (A.M.)	$\bar{x}_1$	$\bar{x}_2$

Then, the mean of the two sets of data X_1 and X_2 together, i.e., the combined mean denoted by $\bar{X}$ is given by

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

Similarly, if we consider ‘k’ sets of data say $X_1, X_2, \dots, X_k$with their respective number of observations $n_1, n_2, \dots n_k$(say) and means $\bar{x}_1, \bar{x}_2, \dots \bar{x}_k$, then the combined mean $\bar{X}$ is given by

$$\boldsymbol{\bar{X} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2 + \cdots + n_k\bar{x}_k}{n_1 + n_2 + \cdots + n_k}}$$

Ex.5. The average daily wage of 150 workers in a factory is ₹720. The average daily wage of 90 male workers is ₹750. Find the average daily wage of female workers.

Sol.

Given,	Male workers	Female workers
No. of workers	$n_1 = 90$	$n_2 = 150 - 90 = 60$
Average daily wage (₹)	$\bar{x}_1 = 750$	$\bar{x}_2 = ?$

Average daily wage of 100 workers $\bar{X} = ₹720$

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

=> 720 = $\frac{90\times750+60\times\bar{x}_2}{90+60}$

=> 108000 = 67500 + $60\bar{x}_2$

=> 108000 – 67500 = $60\bar{x}_2$

=> 40500 = $60\bar{x}_2$

=> $\bar{x}_2 = \frac{40500}{60} = 675$

Hence, the average daily wage of female workers is ₹675

Ex.6. Average marks obtained in a certain subject by boys was 85. Average marks obtained in the same subject by the girls was 81. The average marks obtained by the students combined was 84. Find the percentage of the students.

Sol.

Given,	Boys	Girls
No. of students	n_1 =?	n_2 =?
Average marks obtained	$\bar{x}_1 = 85$	$\bar{x}_2 = 81$

Average marks obtained by all the students was $\bar{x} = 84$

So,

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

=> 84 =$\frac{n_1\times85+n_2\times81}{n_1+n_2}$

=> $84 \times n_1 + 84 \times n_2 = 85n_1 + 81n_2$

=> $84n_1 + 84n_2 = 85n_1 + 81n_2$

=> $84n_1 - 85n_1 = 81n_2 - 84n_2$

=> $-n_1 = -3n_2$

=> $n_1 = 3n_2$

=> $\frac{n_1}{n_2} = \frac{3}{1}$

=> $n_1 : n_2 = 3 : 1$

Therefore, the percentage of boys = $\frac{3}{3+1} \times 100\% = 75\%$

And the percentage of girls = $\frac{1}{3+1} \times 100\% = 25\%$

Ex.7. If the average height of 30 men is 158 cm and the average height of another group of 40 men is 162 cm, find the average height of the combined group.

Hints: $n_1 = 30, n_2 = 40, \bar{x}_1 = 158, \bar{x}_2 = 162$, find $\bar{X}$.

Merits or advantages of A.M.

(i) It is rigidly defined.

(ii) It is easy to understand and easy to calculate.

(iii)Its calculation is based on all the values of the variable.

(iv)It is suitable for further mathematical treatments

(v) It is less effected by sampling fluctuations i.e, it is stable.

Demerits or disadvantages of A.M.

(i) It cannot be calculated in case of open-end class distribution.

(ii) It cannot be determined graphically.

(iii)It is effected by the extreme values.

(iv)This average cannot be calculated in qualitative data.

Uses of A.M.

(i) A.M. is used to find average marks secured by the students, average income of employees in an industry, average profit and so on.

(ii) A.M. is considered to be the best average due to its mathematical properties. So, it is used in computation of various other statistical measures such as standard deviation, coefficient of skewness, correlation and regression etc.

Geometric Mean (G.M.)

In case of simple or ungrouped data, if $X_1, X_2, X_3, \ldots, X_n$be n positive numbers of a variable X, then the G.M. is defined as $G.M. = (X_1 \times X_2 \times X_3 \times \ldots \times X_n)^{\frac{1}{n}}$

Or $G.M. = antilog(\frac{\sum logX}{n})$

In case of grouped data if $f_1, f_2, f_3, \dots, f_n$ be the corresponding frequencies of the n positive values say $X_1, X_2, X_3, \dots, X_n$,then

$$G.M. = \left(X_1^{f_1} \times X_2^{f_2} \times X_3^{f_3} \times \dots \times X_n^{f_n}\right)^{\frac{1}{\Sigma f}}$$

Or $G.M. = antilog\left(\frac{\Sigma flogX}{\Sigma f}\right)$

Merits / Advantages of G.M.

1) It is rigidly defined.
2) It is based on all the observations.
3) It is capable of further mathematical treatment.
4) It is less effected by extreme values

Demerits / Disadvantages of G.M.

1) It is not easy to understand and easy to calculate.
2) If any one of the values is zero, the value of G.M. is also zero. That is, it ignores the other values.
3) If any of the values is negative, the value of G.M may be imaginary.

Uses of G.M.

1) It is used to determine the average of interest, rates, ratios of data.
2) It is used as a good average in the construction of an ideal Index Number.

Harmonic Mean (H.M.)

In case of simple or ungrouped data, if $X_1, X_2, X_3, \dots, X_n$be the non-zero values of a variable X, then H.M. = $\frac{Number\ of\ values}{sum\ of\ the\ reciprocals\ of\ the\ values}$

Thus, H.M. = $\frac{n}{\frac{1}{X_1}+\frac{1}{X_2}+\frac{1}{X_3}+\cdots\frac{1}{X_n}} = \frac{n}{\Sigma\frac{1}{X}}$

In case of grouped data, if $f_1, f_2, f_3, \dots, f_n$ be the corresponding frequencies of the n non-zero values say$X_1, X_2, X_3, \dots, X_n$, then H.M. = $\frac{f_1+f_2+f_3+\cdots+f_n}{\frac{f_1}{X_1}+\frac{f_2}{X_2}+\frac{f_3}{X_3}+\cdots+\frac{f_n}{X_n}} = \frac{\Sigma f}{\Sigma\frac{f}{X}}$

Ex.8. Determine H.M. of the following numbers

46.1, 121, 127, 202

Sol.

$H.M. = \frac{n}{\sum \frac{1}{X}}$

$= \frac{4}{0.0217+0.0083+0.0079+0.0050}$

$= \frac{1}{0.0429} = 93.24$

Ex.9. Find the Harmonic Mean of the following distribution:

Class	**: 10 – 14**	**15 – 19**	**20 – 24**	**25 – 29**	**30 – 34**
Frequency:	**4**	**6**	**8**	**2**	**1**

Sol.

Class	**Frequency f**	**Mid value** **X**	$\boldsymbol{\frac{f}{X}}$
10 – 14	4	12	$\frac{4}{12} = 0.3333$
15 – 19	6	17	$\frac{6}{17} = 0.3529$
20 – 24	8	22	$\frac{8}{22} = 0.3636$
25 – 19	2	27	$\frac{2}{27} = 0.0741$
30 – 34	1	32	$\frac{1}{32} = 0.0313$
Total	$\sum f$=21		$\sum \frac{f}{X}$=1.1552

H.M. = $\frac{\sum f}{\sum \frac{f}{X}} = \frac{21}{1.1552} = 18.18$

Application problems of Harmonic Mean

Ex.10. A motor car covers the first 30 km at the speed of 15 km per hour, the second 30 km at the speed of 20 km per hour and the last 30 km at the speed of 25 km per hour. Find the average speed of the car.

Sol.

Let, X denotes the speed in km per hour and f denotes distance covered in km.

X:	15	20	25
f:	30	30	30

Therefore, the average speed is

$$\frac{\sum f}{\sum \frac{f}{x}} = \frac{30+30+30}{\frac{30}{15}+\frac{30}{20}+\frac{30}{25}}$$

$$= \frac{90}{2+1.5+1.2}$$

$$= \frac{90}{4.7}$$

$$= 19.15\ km\ per\ hour$$

Ex.11. A man travelled from one place to another place at the rate of 20 km per hour and returned at the rate of 30 km per hour. Find the average speed of the whole journey.

Sol.

Let X denotes the speed in km per hour

X: 20 30

The average speed is

$$\frac{n}{\sum \frac{1}{X}} = \frac{2}{\frac{1}{20}+\frac{1}{30}}$$

$$= \frac{2}{\frac{3+2}{60}}$$

$$= \frac{2\times 60}{5} = 24\ km\ per\ hour$$

Merits / Advantages of H.M.

(i) It is rigidly defined.

(ii) It is based on all the values of the distribution.

(iii)It is suitable for further mathematical treatment.

(iv)It is suitable in case of series having wide dispersion.

Demerits / Disadvantages of H.M.

(i) It is not easy to understand and the easy to calculate.

(ii) If any value is zero, the value of H.M. is also zero, so it ignores the other values.

(iii)I gives more weight to the smaller value.

Uses of H.M.

It is used to determine the average speed of a moving body travelled in equal interval of time.

Relationship between A.M., G.M. and H.M.

(i) $A.M. \geq G.M. \geq H.M.$

(ii) G.M.=$\sqrt{A.M.\times H.M.}$

Or $(G.M.)^2 = A.M.\times H.M.$

Ex.12. If GM = 6, AM = 8, then HM = ?

Ans: We have

$(G.M.)^2 = A.M.\times H.M.$

=> $6^2 = 8 \times HM$

=> $36 = 8 \times HM$

=> HM $= \frac{36}{8} = 4.25$

Positional Averages

- ✓ Median
- ✓ Mode

Median: It is that value of a variate which divides the distribution into two equal parts, when the values are arranged according to their magnitudes (ascending order or descending order).

Calculations of median:

In case of ungrouped data,

$Median = the\ value\ of\ \left(\frac{\boldsymbol{N+1}}{\boldsymbol{2}}\right) th\ observation, where, N = \sum f\ is\ the\ number\ of\ observations.$

In case of grouped data,

$$Median = \boldsymbol{L} + \frac{\frac{\boldsymbol{N}}{2} - \boldsymbol{C}}{\boldsymbol{f}} \times \boldsymbol{h}$$

Where, L = lower limit of the median class.

h = width of the median class

= Upper median class – Lower median class

C = cumulative frequency (CF) of the preceding of the median class.

f = corresponding frequency of the median class.

N = total frequency i.e, $\sum f$

Ex.13. Determine median of the following series:

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

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

Sol.

(i) Arranging the values of the series in ascending order we get,

70, 72, 73, **75**, 77, 78. 79

The no. of terms N = 7

Therefore,

Median = the value of $\left(\frac{N+1}{2}\right) th$ observation

= the value of $\left(\frac{7+1}{2}\right) th$

= 4th observation

= 75

(ii) Arranging the values of the series in ascending order we get,

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

The number of observations N = 8

Therefore,

Median = the value of $\left(\frac{N+1}{2}\right) th$ observation

= the value of $\left(\frac{8+1}{2}\right) th$ observation

= the value of 4.5th observation

= the value of $\frac{4th\ observation+5th\ observation}{2}$

= $\frac{68+70}{2}$

= 69

Ex.14. Determine Median for the following distribution

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

Sol.

Wages(₹)	**No. of workers (f)**	**Cumulative frequency (CF)**
20	8	8 (1st – 8th)obs
21	10	8+ 10 = 18 (9th – 18th)obs
22	11	18 + 11 = 29 (19th – 29th)obs
23	16	29 + 16 = 45 (30th – 45th) obs
24	20	45 + 20 = 65 (46th – 65th) obs
25	25	65 + 25 = 90 (66th – 90th) obs
26	19	90 + 19 = 109 (91st – 109th) obs
27	9	109 + 9 = 118

		(110th – 118th) obs
28	6	118 + 6 = 124 = N (119th – 124th) obs
Total	$\sum f$=124	

∴ Median = the value of $\left(\frac{N+1}{2}\right)th$ observation

= the value of $\left(\frac{124+1}{2}\right)th$ observation

= the value of 62.5th observation

= the value of $\frac{62th\ observation+63th\ observation}{2}$

= $\frac{24+24}{2}$ = 24

Ex.15. Find median marks obtained by the students

Marks obtained:	**5**	**10**	**15**	**20**	**25**
No. of students :	**3**	**4**	**2**	**5**	**4**

Sol.

Marks obtained	**No. of students (f)**	**CF**
5	3	3 (1st – 3rd) obs
10	4	3 + 4 = 7 (4th – 7th) obs
15	2	7 + 2 = 9 (8th – 9th) obs
20	5	9 + 5 = 14 (10th – 14th) obs
25	4	4 + 4 = 18 = N (15th – 18th) obs

Total	$\sum f$= 18	

Median = the value of $\left(\frac{N+1}{2}\right) th$ observation

= the value of $\left(\frac{18+1}{2}\right) th = 9.5th$ observation

= the value of $\frac{9th\ observation + 10th\ observation}{2}$

= $\frac{15+20}{2}$ = 17.5 marks

Ex.16. Determine median for the following distribution

Daily wages (₹): 50 – 55 55 – 60 60 – 65 65 – 70 70 – 75 75 – 80 80 – 85

No. of workers: 6 10 22 30 16 12 7

Sol.

Daily wages (₹)	**No. of students**	**CF**
50 – 55	6	6
55 – 60	10	16
60 – 65	22	38 C
L 65 – 70	**30 f**	**68**
70 – 75	16	84
75 – 80	12	96
80 – 85	7	103 = N
Total	$\sum f$ = 103	

$$\frac{N}{2} = \frac{103}{2} = 51.8$$

The value of CF just greater than 51.8 is 68

So, the median class is 65 – 70

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

$= 65 + \frac{51.8-38}{30} \times 5$

= 65 + 0.46 = ₹65.46

If a distribution has inclusive class interval, to calculate median, we have to convert the inclusive class intervals to exclusive intervals by using class boundaries.

Ex.17. Calculate Median from the following distribution

Class interval: 130 – 134 135 – 139 140 – 144 145 – 149 150 – 154 155 – 159 160 – 164

Frequency: 5 15 28 24 17 10 1

Sol.

Class boundary	$\boldsymbol{f}$	$\boldsymbol{CF}$
129.5 – 134.5	5	5
134.5 – 139.5	15	20
139.5 – 144.5	28	48 C
L 144.5 – 149.5	**f 24**	**72**
149.5 – 154.5	17	89
154.5 – 159.5	10	99
159.5 – 164.5	1	100 = N
Total	$\sum f$=100	

$\frac{N}{2} = \frac{100}{2} = 50$

The value of CF just greater than 50 is 72

So, the median class is 144.5 – 149.5

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

$= 144.5 + \frac{50-48}{24} \times 5$

= 144.5 + 0.42 = 144.92

Merits / Advantages of Median

(i) It is easy to understand and simple to calculate.

(ii) It is not effected by the extreme values.

(iii)It can be determined graphically

(iv) It can be determined in case of open-end class distribution.

Demerits / Disadvantages of Median:

(i) Data should be arranged in ascending or descending order.

(ii) Its calculation is not based on all the values of the distribution.

(iii)It is not suitable for further algebraic treatments.

Uses:

(i) It is used to determine the average of distribution having open-end class intervals.

(ii) Median is used to study the income distribution.

Mode: Mode is that value of the variate which occurs the maximum number of times i.e., the value with the maximum frequency.

Ex.18. Find the mode of the distribution of values 5, 9, 7, 7, 5, 9, 6, 7, 5, 4, 3, 4, 1, 5

Sol.

X	f
1	1
3	1
4	2
5	**4**
6	1
7	3
9	2

Therefore, mode = corresponding value of the maximum frequency 4 = 5

In case of grouped data,

$$\boldsymbol{Mode = L + \frac{f_m - f_0}{2f_m - f_0 - f_1} \times h}$$

Where, L = lower limit of the modal class.

h = width of the modal class.

f_m= corresponding frequency of the modal class i.e, the maximum frequency.

f_0 = corresponding frequency of the preceding of the modal class.

f_1 = corresponding frequency of the succeeding of the modal class.

Ex.19. Determine mode for the following distribution

Daily wages (₹):	**50 – 55**	**55 – 60**	**60 – 65**	**65 – 70**	**70 – 75**	**75 – 80**	**80 – 85**
No. of workers:	**6**	**10**	**22**	**30**	**16**	**12**	**15**

Sol.

Daily wages (₹)	**No. of workers (f)**
50 – 55	6
55 – 60	10
60 – 65	22 f_0

L 65 – 70	**30 f_m**
70 – 75	16 f_1
75 – 80	12
80 – 85	15

The maximum frequency f_m = 30

∴ The modal class is 65 – 70

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

=> Mode = $65 + \frac{30-22}{2\times30-22-16} \times 5$

= 65 + 1.82 = ₹66.82

If a distribution has inclusive class interval, to calculate mode, we have to convert the inclusive class intervals to exclusive intervals by using class boundaries.

Ex.20. Calculate Mode from the following distribution

Class interval:	**130 – 134**	**135 – 139**	**140 – 144**	**145 – 149**	**150 – 154**	**155 – 159**	**160 – 164**
Frequency:	**5**	**15**	**28**	**24**	**17**	**10**	**1**

Sol.

Class boundary	***f***
129.5 – 134.5	5
134.5 – 139.5	15 f_0
L 139.5 – 144.5	**28 f_m**
144.5 – 149.5	24 f_1
149.5 – 154.5	17
154.5 – 159.5	10
159.5 – 164.5	1

∴Mode = $L + \frac{f_m - f_0}{2f_m - f_0 - f_1} \times h$

= $139.5 + \frac{28-15}{2\times28-15-24} \times 5$

= 139.5 + 3.82 = 143.32

Merits / Advantages of Mode:

(i) It is easy to understand and simple to calculate.

(ii) It is not affected by extreme values.

(iii) It can be determined graphically.

Demerits / Disadvantages of Mode:

(i) It is not based on all the observations.

(ii) It is not suitable for further mathematical treatment.

(iii)Some distribution may have two or more mode values.

Uses of Mode:

(i) Mode is used to determine the ideal measure of manufactured / business items like the sizes of shoes, garments etc.

(ii) It is also used in weather forecast.

Empirical Relationship between Mean, Median and Mode:

In a symmetric distribution, mean, median and mode are approximately equal,

i.e., **mean = median = mode**

In a moderately asymmetrical (skewed) distribution,

Mean – Mode = 3 × (Mean – Median)

Or **Mode = 3 × Median – 2 × Mean**

Ex. 21. In a moderately skewed distribution, if median = 35, mode = 31.5, find mean.

Sol.

We have,

Mode = 3median – 2mean

Or 31.5 = 3×35 - 2×mean

Or 31.5 = 105 - 2×mean

Or 2×mean = 105 – 31.5

Or 2×mean = 73.5

Or mean = $\frac{73.5}{2}$

Or mean = 36.75

Ex.22. Find mode

Class	**:10 – 20**	**20 – 30**	**30 – 40**	**40 – 50**	**50 – 60**
Frequency:	**5**	**8**	**12**	**16**	**18**

Sol.

By the question, $f_m = 18, f_0 = 16$, but there is no f_1. So, in this case, mode is ill defined. To calculate mode, empirical relations for a moderately asymmetrical distribution is used.

Class	**Frequency f**	$\boldsymbol{CF}$	**Mid Value X**	$\boldsymbol{fX}$
10 – 20	5	5	15	75
20 – 30	8	13	25	200
30 – 40	12	**25 C**	35	420
L 40 – 50	**16 f**	**41**	45	720
50 – 60	18	59 = N	55	990
Total	$\sum f$**= 59**			$\sum fX$**= 2405**

Mean = $\frac{\sum fX}{\sum f}$

$= \frac{2405}{59} = 40.76$

$\frac{N}{2} = \frac{59}{2} = 29.5$

The value of CF just greater than 29.5 is 41.

So, the median class is 40 – 50

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

$= 40 + \frac{29.5 - 25}{16} \times 10$

= 40 + 2.81

= 42.81

By using empirical relation,

$Mode = 3 \times Median - 2 \times Mean$

Or mode $= 3 \times 42.81 - 2 \times 40.76$

Or mode $= 46.91$

Determination of Median and Mode by using Graph

To determine median, two ogives (cumulative frequency curves) – less than and greater than are drawn. The value on the horizontal line (X-axis) at which the perpendicular line passing through the point of intersection of two ogives, tersects is the median value.

To determine the mode value, histograms are drawn. The value on the horizontal line (X-axis) at which the perpendicular line passing through the point of intersection between the two lines connecting the corner points of the two adjacent rectangles of the longest rectangle of the histogram.

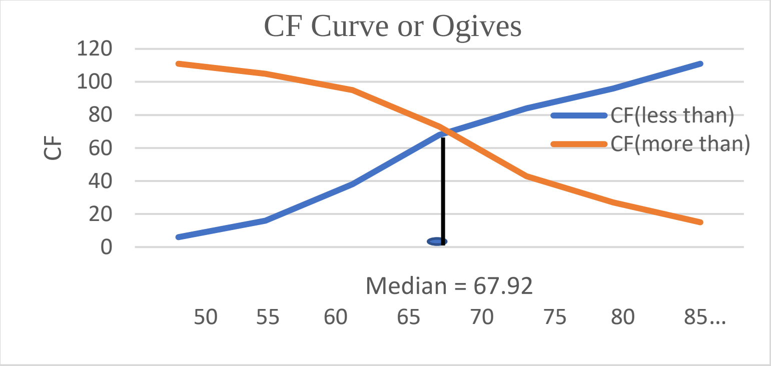
Ex.23. Determine median and mode for the following distribution

Daily wages (₹):	**50 – 55**	**55 – 60**	**60 – 65**	**65 – 70**	**70 – 75**	**75 – 80**	**80 – 85**
No. of workers:	**6**	**10**	**22**	**30**	**16**	**12**	**15**

Sol.

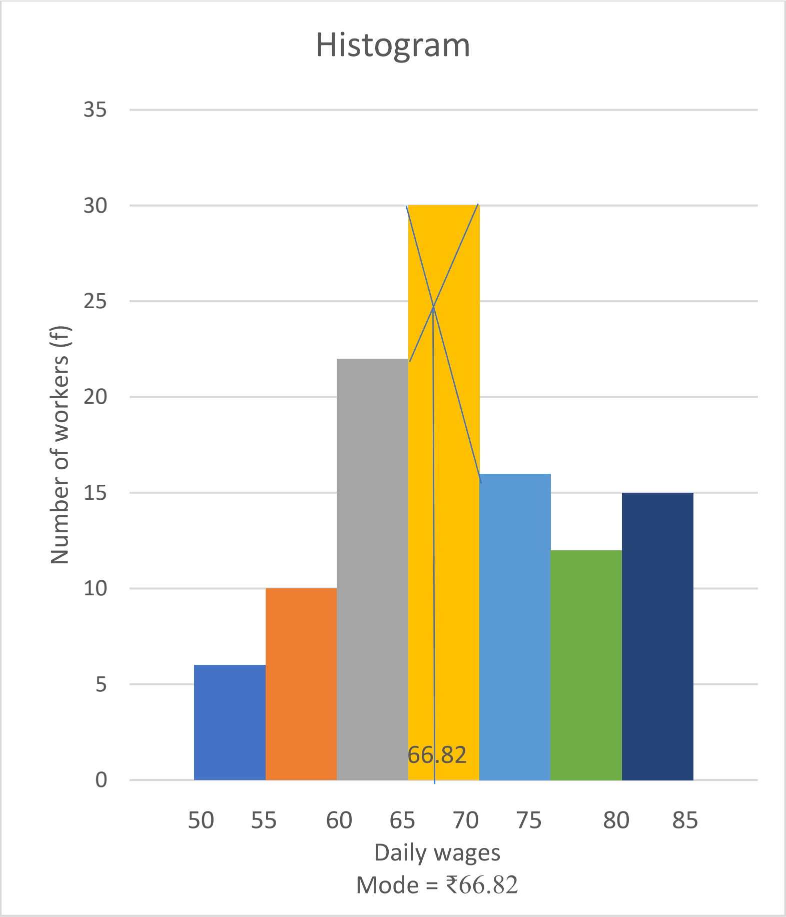
Determination of Median

Daily wages	**No. of Workers (f)**	**Upper limit**	**CF(less than)**	**Lower limit**	**CF(more than)**
50 – 55	6	55	6	50	111
55 – 60	10	60	6+10=16	55	111 – 6 = 105
60 – 65	22	65	16+22=38	60	105 – 10 = 95
65 – 70	30	70	38+30=68	65	95 – 22 = 73
70 – 75	16	75	68+16=84	70	73 – 30 = 43
75 – 80	12	80	84+12=96	75	43 – 16 = 27
80 – 85	15	85	96+15=111 = N	80	27 – 12 = 15



Determination of Mode:

Daily wages	**No. of Workers (f)**
50 – 55	6
55 – 60	10
60 – 65	22
65 – 70	30
70 – 75	16
75 – 80	12
80 – 85	15



1.1.Measures of Dispersion (Variation)

Literally, dispersion means scatteredness. We study dispersion to have an idea about the homogeneity or heterogeneity of the distribution.

“The variation or scattering or deviation of the different values of a variable from their average, is known as dispersion”

For example, let us consider the following two sets of data-

(i) 30, 40, 50, 60, 70 mean = 50

(ii) 100, 80, 20, 40, 10 mean = 50

In (i) the data are less variation i.e. more consistent whereas, in (ii) the data are more variation, i.e, less consistent. In measures of central tendency, we cannot study the consistency or homogeneity of data. To overcome this problem, the measures of dispersion are applied.

Essential qualities of a good measure of dispersion / Characteristics of an Ideal measure

of dispersion:

(i) It should be rigidly defined.

(ii) It should be easy to understand and easy to calculate.

(iii)It should be based on all the observations.

(iv)It should be suitable for further algebraic treatments.

(v) It should be least effected by the fluctuations of the sampling.

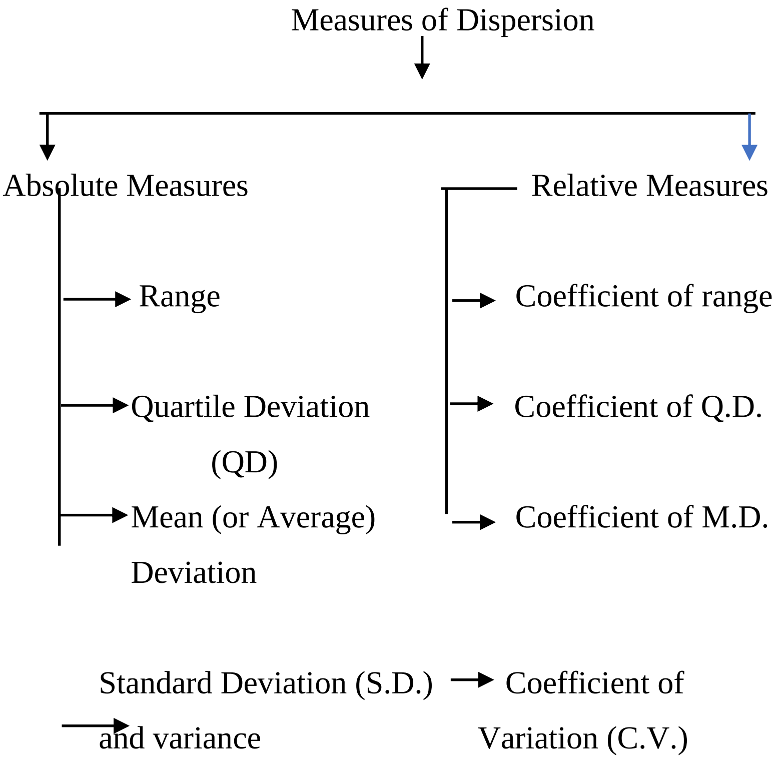
Purposes / Objectives / Significance of measuring dispersion:

(i) **To test the reliability of an average:** Measures of dispersion are used to test to what extent an average represents the characteristics of the values of the variable.

(ii) **To control the variability:** measures of dispersion helps to identify the nature and causes of variation. Such information is useful in controlling the variations.

(iii) **To compare two or more sets of data with respect to their variability:** Measures of dispersion help in the comparison of two or more sets of data with respect to their uniformity or consistency.

(iv) **To facilitate the use of other statistical tools:** Measures of dispersion facilitate the use of other statistical techniques like skewness, kurtosis, correlation, regression etc.



Absolute Measures of Dispersion:

Absolute measures of dispersion are described by a number or value to represent the amount of variation or differences among the values. Such a number or value is expressed in the same unit of measurement as the set of values in the data such as rupees, inches, feet, kilograms or tons. Such measures help in comparing two or more sets of data in terms of absolute magnitudes of variation, provided the various values are expressed in the same unit of measurement and have almost the same average value.

Relative Measures of Dispersion:

A relative measure of dispersion is calculated as the percentage or ratio or the coefficient of absolute measure of dispersion. So, it is also known as the **coefficient of dispersion**. This measure is a pure number i,e, it is free from any unit. Such measures help in comparing two or more sets of data in terms of relative magnitude of variation and the values are expressed in terms of same unit.

Range:

The range is the simplest measure of dispersion. The range is defined to be the difference between the largest and the smallest observed values in a set of values.

Thus, **Range = L – S**, where, L is the largest value and S is the smallest value.

The relative measure of range, called the coefficient of range, is obtained by applying the following formula:

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

Ex.24. Find range and its coefficient from the following data:

Weight (Kg) : 40, 51, 47, 39, 60, 48, 64, 61, 57

Sol.

Range = L – S = 64 – 39 = 25 Kgs

Coefficient of range = $\frac{L-S}{L+S} = \frac{64-39}{64+39} = \frac{25}{103} = 0.2427$

Ex.25. Find Range and its coefficient from the following frequency distribution:

Daily Wages (₹):	**6 – 10**	**10 -14**	**14 – 18**	**18 – 22**
# workers :	**2**	**6**	**10**	**4**

Sol.

Range = L – S = 22 – 6 = ₹16

Coefficient of Range = $\frac{L-S}{L+S} = \frac{22-6}{22+6} = \frac{16}{28} = 0.5714$

Merits / Advantages of Range:

(i) It is easy to understand and easy to calculate.

(ii) It is quite useful in cases where the purpose is only to find the extent of extreme variation, such as industrial quality control, temperature etc.

Demerits / Disadvantages of Range:

(i) It is very much affected by the extreme values.

(ii) It is not based on all the values of variable.

(iii)It cannot be calculated from open-end class distribution.

Uses of Range:

(i) **Fluctuation in share prices:** the range is useful in the study of variation of share prices and other commodities that are very sensitive to price changes from one period to another.

(ii) **Quality control:** It is widely used in industrial quality control, which is exercised by preparing suitable control charts based on range.

(iii)**Weather forecast:** The concept of range is used to determine the difference between the maximum and the minimum temperatures or rainfall by meteorological departments to announce for the knowledge of the general public.

Quartiles: Quartiles are those values of the variate which divide the distribution into four equal parts when the values are arranged according to their magnitudes. There are three quartiles viz., **first or lower quartile or Q_1, second quartile or Q_2 or Median and third or higher quartile or Q_3.**

In case of simple or discrete data,

$Q_k = the\ value\ of\ \frac{k(N+1)}{4} th\ observation.$,

where, N = $\sum f$ = the number of observations.

and k = 1, 2, 3.

In case of grouped data,

$$\boldsymbol{Q_k = L + \frac{\frac{kN}{4} - C}{f} \times h}$$

Where, k = 1, 2, 3

L = lower limit of the class where the kth quartile exists.

h = width of the class where, the kth quartile exists.

C = cumulative frequency of the preceding of the class, where, the kth quartile exists.

f = the corresponding frequency of the class, where, the kth quartile exists.

N = the total frequency i.e, $\sum f$

Quartile Deviation (Q.D.)

The half of the difference between the higher and the lower quartiles, is called quartile deviation. In other words, the semi quartile range is called quartile deviation.

Thus, **Q.D. =** $\boldsymbol{\frac{Q_3 - Q_1}{2}}$ = **semi interquartile range**

Where,$Q_3 - Q_1$ is known as inter quartile range.

The relative measure of Q.D. is given by

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

Ex.26. Find the quartile deviation and its coefficient from the following data

19, 27, 24, 39, 57, 44, 56, 50, 59, 67, 62, 42, 47, 60, 26, 44, 57, 51, 59, 45

Sol.

Arranging the data in ascending order, we have

19, 24, 26, 27, **34, 39**, 42, **44, 45**, 47, 50, **51, 56**, 57, **57, 59**, 59, 60, 62, 67

Q_1= the value of $\frac{20+1}{4} th$ observation

= the value of 5.25th observation.

= the value of 5th observation

+ 0.25 × (6th observation – 5th observation)

= 34 + 0.25 × (39 – 34)

= 34 + 1.25 = 35.25

$Q_3 =$ the value of $\frac{3\times(20+1)}{4} th$ observation

= the value of 15.75th observation

= the value of 15th observation

+ 0.75 × (16th observation – 15th observation)

= 57 + 0.75 × (59 – 57)

= 57 + 1.5 = 58.5

Therefore, quartile deviation is

$QD = \frac{Q_3 - Q_1}{2}$

$= \frac{58.5-35.25}{2} = 11.625$

Coefficient of QD $= \frac{Q_3-Q_1}{Q_3+Q_1} = \frac{58.5-35.25}{58.5+35.25} = 0.248$

Ex.27. Determine the quartile deviation and its relative measure from the following distribution.

Marks	**: 10**	**20**	**30**	**40**	**50**	**60**	**70**	**80**
Number of students:	**5**	**25**	**40**	**70**	**90**	**40**	**20**	**10**

Sol.

Marks	Number of students (f)	Cumulative frequency (CF)
10	5	5 (1st – 5th) obs
20	25	30 (6th – 30th) obs
30	40	70 (31st – 70th) obs
40	70	**140 (71st – 140th) obs**
50	90	**230 (141st – 230th) obs**
60	40	270 (231st – 270th) obs
70	20	290 (271st – 290th) obs
80	10	300 = N (291st – 300th) obs

Q_1= the value of $\frac{300+1}{4} th$ observation

= the value of 75.25th observation

= the value of 75th observation + 0.25 × (76th observation – 75th observation)

= 40 + 0.25 × (40 – 40) = 40

Q_3= the value of $\frac{3\times(300+1)}{4} th$ observation

= the value of 225.75th observation.

= the value of 225th observation + 0.75 × (226th observation – 225th observation)

= 50 + 0.75×(50 – 50) = 50

Therefore, the $QD = \frac{Q_3 - Q_1}{2}$

$= \frac{50-40}{2} = 5$

The coefficient of QD = $\frac{Q_3-Q_1}{Q_3+Q_1}$

$= \frac{50-40}{50+40} = 0.1111$

Ex.28. A survey of domestic consumption of electricity gave the following distribution of the units consumed. Compute the quartile deviation and its coefficient.

No. of units : below 200 200 – 400 400 – 600 600 – 800 800 – 1000 1000 – 1200 1200 – 1400 1400 & above

No. of consumers: 9 18 27 32 45 38 20 11

Sol.

No. of units	No. of consumers (f)	Cumulative Frequency
Below 200	9	9
200 – 400	18	27 C
L400 – 600	**27 f**	**54**
600 – 800	32	86
800 – 1000	45	131 C
L1000 – 1200	**38 f**	**169**
1200 – 1400	20	189
Above 1400	11	200 = N
Total	$\sum f$= 200	

$\frac{N}{4} = \frac{200}{4} = 50$

The value of CF just greater than 50 is 54

Therefore, Q_1 lies in 400 – 600

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

$$= 400 + \frac{50-27}{27} \times 200$$

$$= 570.37$$

$\frac{3N}{4} = \frac{3 \times 200}{4} = 150$

The value of CF just greater than 150 is 169

Therefore, Q_3 exists in 1000 – 1200

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

$$= 1000 + \frac{150-131}{38} \times 200$$

= 1100

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

$= \frac{1100-570.37}{2} = 264.815$ units

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

$$= \frac{1100-570.37}{1100+570.37}$$

$$= \frac{529.63}{1670.37} = 0.3171$$

Merits / Advantages of Q.D.

(i) It is easy to understand and easy to calculate.

(ii) Since, its calculation is based on 50% of the observation, so it is superior to range.

(iii) It is not affected by the extreme values.

(iv) It can be computed in case of open-end class distribution.

Demerits / Disadvantages of Q.D.

(i) It is not based on all the values of the observation.

(ii) It is not suitable for further algebraic treatment.

Uses of Q.D.:

(i) Since, the calculation of Q.D. is not affected by the presence of extreme values, so, it is used in open-end class distribution.

(ii) It is one of a good measure of dispersion used in Descriptive Statistics.

Mean (Average) Deviation (MD):

In case $X_1, X_2, \dots, X_n$be the values of a variable X and $\bar{X}$, the mean or median or mode of the distribution, then the MD is defined as **MD =** $\boldsymbol{\frac{\sum |X-\bar{X}|}{n}}$

In case of grouped data, if $f_1, f_2, \dots, f_n$ be the corresponding frequencies of the values $X_1, X_2, \dots X_n$ of the variable X and $\bar{X}$, the mean or median or mode, then MD is given by

MD $= \boldsymbol{\frac{\sum f|X-\bar{X}|}{\sum f}}$

The relative measure of MD is defined as **Coefficient of MD** $= \boldsymbol{\frac{MD}{\bar{X}}}$

Ex.29. Determine Mean Deviation from mean and its coefficient of the following distribution

Weight (Kg) : 40, 51, 47, 39, 60, 48, 64, 61, 57

Sol.

Weight (Kg) X	$\boldsymbol{X-\bar{X}=X-51.89}$	$\boldsymbol{\lvert X-\bar{X}\rvert}$
40	-11.89	11.89
51	-0.89	0.89
47	-4.89	4.89
39	-12.89	12.89
60	8.11	8.11
48	-3.89	3.89
64	12.11	12.11
61	9.11	9.11

57	5.11	5.11
$\sum X$=467	**-0.01= 0**	**$\sum \lvert X-\bar{X}\rvert$=68.89**

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

$= \frac{467}{9}$

$= 51.89$

Therefore, the MD from mean $= \frac{\sum |X-\bar{X}|}{n}$

$= \frac{68.89}{9}$ $= 7.654$ kg

The coefficient of MD = $\frac{MD}{\bar{X}}$

$= \frac{7.654}{51.89} = 0.1475$

Ex.30. Calculate mean deviation from mean for the following data. Also find the coefficient of mean deviation

Class interval:0 – 4 4 – 8 8 – 12 12 – 16 16 – 20

Frequency : 4 6 8 5 2

Sol.

C. I.	f	**Mid value x**	fx	**|x - $\bar{x}$|**	$f\lvert x-\bar{x}\rvert$
0 – 4	4	2	8	7.2	28.8
4 – 8	6	6	36	3.2	19.2
8 – 12	8	10	80	0.8	6.4
12 – 16	5	14	70	4.8	24.0
16 – 20	2	18	36	8.8	17.6
Total	$\sum f$ **= 25**		$\sum fx$ **=230**		$\sum f\lvert x-\bar{x}\rvert$ **= 96**

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

$= \frac{230}{25} = 9.2$

Mean deviation from mean $= \frac{\sum f|X-\bar{X}|}{\sum f}$

$= \frac{96}{25} = 3.84$

Coefficient of MD = $\frac{MD}{\bar{X}}$

$= \frac{3.84}{9.2} = 0.4174$

Merits / Advantages of MD

(i) It is based on all the observations, if it is taken from mean.

(ii) It is less affected by the extreme values, if it is taken from median.

(iii)It is a good measure for comparing the variability among two or more distribution.

Demerits / Disadvantages of MD

(i) It ignores the negative sign.

(ii) MD from mode is not considered to be a good measure.

(iii)It is not as popular as the standard deviation

Uses of MD

Due to its simplicity in computation, it is used in studying economic and business problems. It is also used to study the variability of income and wealth distribution.

Self study: Ex. 31. Calculate MD and its coefficient from the following distribution

Daily Wages (₹):	**6 – 10**	**0 – 14**	**14 – 18**	**18 – 22**
# workers :	**2**	**6**	**10**	**4**

[Answers: MD = ₹2.84, Coefficient of MD = 0.1905]

Standard Deviation and Variance

In case of simple or ungrouped data, if $X_1, X_2, \dots X_n$ be the values of a variable X and $\bar{X}$, the mean of X, then the standard deviation of X, denoted by σ, is defined as

$$\boldsymbol{\sigma = \sqrt{\frac{\sum(X-\bar{X})^2}{n}}}$$

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

In case of grouped data, if $f_1, f_2, \dots, f_n$ be the corresponding frequencies of the values $X_1, X_2, \dots, X_n$ of the variable X and $\bar{X}$, the mean of X, then the standard deviation of X is given by $\boldsymbol{\sigma = \sqrt{\frac{\sum f(X-\bar{X})^2}{\sum f}}}$ or $\boldsymbol{\sigma = \sqrt{\frac{\sum fx^2}{\sum f} - \left(\frac{\sum fx}{\sum f}\right)^2}}$

The square of the standard deviation is called the **variance**. If σ is the standard deviation, then the variance $=\sigma^2$. Thus, standard deviation $= \sqrt{variance}$

Coefficient of variation (CV):

It is the best relative measure of the dispersion. It is expressed as the percentage of the ratio of the standard deviation to the mean. Thus,

$$\boldsymbol{CV = \frac{Standard\ deviation}{mean} \times 100\%}$$

Or $\boldsymbol{CV = \frac{\sigma}{\bar{X}} \times 100\%}$

The value with less CV value is more consistent or more homogeneous or less variation from the average.

Ex.32. Find the variance, standard deviation and the coefficient of variation of the following data.

Weight (Kg): 40, 51, 47, 39, 60, 48, 64, 61, 57

Sol.

Weight (Kg) X	$\boldsymbol{X^2}$
40	1600
51	2601
47	2209
39	1521
60	3600
48	2304
64	4096
61	3721
57	3249
$\boldsymbol{\sum X}$**=420**	$\boldsymbol{\sum X^2}$**=24901**

Here, n = 9

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

$$= \frac{24901}{9} - \left(\frac{420}{9}\right)^2$$

$= 2766.7778 - 2177.7778$ = 589 (kg)^2

∴Standard deviation $\sigma = \sqrt{variance}$

$$= \sqrt{589} = 24.27 \text{ kg}$$

Mean $\bar{X} = \frac{\sum X}{n} = \frac{420}{9} = 46.67$ kg

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

$$= \frac{24.27}{46.67} \times 100\% = 52\%$$

Ex.33. Calculate Standard deviation, variance and coefficient of variation from the following distribution.

Daily Wages (₹):	**6 – 10**	**10 – 14**	**14 – 18**	**18-22**
# workers :	**2**	**6**	**10**	**4**

Sol.

Daily Wages(₹)	**# workers f**	**Mid value X**	fX	fX^2
6 – 10	2	8	16	128
10 – 14	6	12	72	864
14 – 18	10	16	160	2560
18 – 22	4	20	80	1600
Total	$\sum f$**=22**		$\sum fX$**=328**	$\sum fX^2$**=5152**

∴Standard deviation

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

$$= \sqrt{\frac{5152}{22} - \left(\frac{328}{22}\right)^2}$$

$$= \sqrt{234.1818 - 14.9091^2}$$

$$= \sqrt{11.9005} = ₹\ 3.45$$

∴ Variance $\sigma^2 = 11.9005(₹)^2$

Mean $\bar{X} = \frac{\sum fX}{\sum f}$

$$= \frac{328}{22} = ₹14.91$$

∴The coefficient of variation (CV) = $\frac{\sigma}{\bar{X}} \times 100\%$

$$= \frac{3.45}{14.91} \times 100\% = 23.14\%$$

Self study:Ex.34. Determine mean, standard deviation, variance and CV of the following distribution.

Class interval:0 – 4 4 – 8 8 – 12 12 – 16 16 – 20

Frequency : 4 6 8 5 2

(Answer: Mean = 9.2, standard deviation = 4.66, variance = 21.76 and CV = 50.65%)

In cumulative frequency distribution, the data are classified in regular class intervals and the calculations of class frequencies are shown in the following example.

Ex.35. Calculate mean, standard deviation and coefficient of variation from the following data:

Age under (years) : 10 20 30 40 50 60 70 80

No. of persons

Dying : 15 30 53 75 100 110 115 125

Sol.

CI	f	X	fX	fX^2
0 – 10	15	5	75	375
10 – 20	30 – 15 = 15	15	225	3375
20 – 30	53 – 30 = 23	25	575	14375
30 – 40	75 – 53 = 22	35	704	24640
40 – 50	100 – 75 = 25	45	1125	50625
50 – 60	110 – 100 = 10	55	550	30250
60 – 70	115 – 110 = 5	65	325	21125
70 – 80	125 – 115 = 10	75	750	56250
Total	$\sum f$**=125**		$\sum fX$**=4329**	$\sum fX^2$**=201015**

∴ Mean $\bar{X} = \frac{\sum fX}{\sum f}$

$$= \frac{4329}{125} \cong 34.63 years$$

∴Standard deviation

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

$$= \sqrt{\frac{201015}{125} - \left(\frac{4329}{125}\right)^2}$$

$$= \sqrt{1608.120 - 1199.375424}$$

$= \sqrt{408.744576} = 20.22$ years

∴ CV $= \frac{\sigma}{\bar{X}} \times 100\%$

$$= \frac{20.22}{34.632} \times 100\% \cong 58.39\%$$

Mathematical properties of Standard Deviation:

(i) The value of a standard deviation cannot be negative i.e, $\sigma \geq 0$

(ii) The standard deviation of a series of equal values is zero

E.g., the standard deviation of the numbers 3, 3, 3, 3 is zero.

(iii)Standard deviation is independent of the change of origin. That is, if the SD of X is σ, then the SD of $X \pm a$ is also σ.

E.g., let,the standard deviation of a series of numbers is 2. If 5 is added to each of the numbers in the series, then the standard deviation of the new set of numbers is also 2.

(iv)Standard deviation is not independent of change of scale. That is, if the SD of X is σ, then the SD of kX is $k\sigma$ and the SD of $\frac{X}{k}$ is $\frac{\sigma}{k}$, where, $k \neq 0$

E.g., let, the standard deviation of a series of numbers is 4. If each of the numbers in the series is divided by the number 2, then the standard deviation of the new series of numbers is $\frac{4}{2} = 2$.

Combined standard deviation (SD)

Let us consider two sets of data X_1 and X_2 such that

	$\boldsymbol{X_1}$	$\boldsymbol{X_2}$
No. of observations	n_1	n_2
Mean	$\bar{X}_1$	$\bar{X}_2$
SD	σ_1	σ_2

Then the combined SD is $\sigma = \sqrt{\frac{n_1(\sigma_1^2+d_1^2)+n_2(\sigma_2^2+d_2^2)}{n_1+n_2}}$

Where, $d_1 = \bar{X}_1 - \bar{X}$ and $d_2 = \bar{X}_2 - \bar{X}, \bar{X} = \frac{n_1\bar{X}_1+n_2\bar{X}_2}{n_1+n_2}$ is the combined mean.

Ex.36. Two groups of workers on the same job show the following results over a long period of time.

	Workers A	**Workers B**
Number of workers	**120**	**150**
Mean time of completing the Job (minutes)	**30**	**25**
Standard deviation	**6**	**8**

(i) Which workers appear to be faster in completing the job?

(ii) Which workers appear to be more consistent in the time they require to complete the job?

(iii)Find the standard deviation of the time of completing the job by all the workers together.

Sol.

Given, $n_1 = 120, n_2 = 150, \bar{X}_1 = 30, \bar{X}_2 = 25, \sigma_1 = 6, \sigma_2 = 8$

(i) Total time of completing the job of workers A

$= n_1 \times \bar{X}_1 = 120 \times 30 = 3600\ minutes$

Total time of completing the job of workers B

$= n_2 \times \bar{X}_2 = 150 \times 25 = 3750\ minutes$

Since, the time of completing of the workers A is less than the time of completing the workers B, so, workers A appear to be faster in completing the job.

(ii) CV of the time of completing of the workers A is $CV_A = \frac{\sigma_1}{\bar{X}_1} \times 100\% = \frac{6}{30} \times 100\%$

$= 20\%$

CV of the time of completing of the workers B is $CV_B = \frac{\sigma_2}{\bar{X}_2} \times 100\% = \frac{8}{25} \times 100\%$

$= 32\%$

Since,$CV_A < CV_B$, so, the workers A appear to be more consistent in the time they require to complete the job.

(iii)Combined mean

$$\bar{X} = \frac{n_1 \times \bar{X}_1 + n_2 \times \bar{X}_2}{n_1 + n_2} = \frac{120 \times 30 + 150 \times 25}{120 + 150} = 27.22 \text{ minutes}$$

$$d_1 = \bar{X}_1 - \bar{X} = 30 - 27.22 = 2.78$$

$$d_2 = \bar{X}_2 - \bar{X} = 25 - 27.22 = -2.22$$

Therefore, the standard deviation of the time of completing the job by all the workers together i.e, the combined standard deviation is

$$\sigma = \sqrt{\frac{n_1(\sigma_1^2 + d_1^2) + n_2(\sigma_2^2 + d_2^2)}{n_1 + n_2}}$$

$$= \sqrt{\frac{120 \times (6^2 + 2.78^2) + 150 \times (8^2 + 2.22^2)}{120 + 130}} = \sqrt{\frac{15586.668}{150}}$$

$$= 10.19\, minutes$$

Merits / Advantages of Standard Deviation:

(i) Standard deviation is rigidly defined.

(ii) It is based on all the values of the variable.

(iii) It is the least affected by sampling variations.

(iv) It is suitable for further algebraic treatment.

(v) Standard deviation is considered to be the best measure of dispersion, because it possesses almost all the requisites of a good measure of dispersion.

Demerits/ Disadvantages of standard deviation:

(i) As compared to other measures, it is neither easy to understand nor simple to calculate.

(ii) It is also affected by extreme values.

Uses of Standard deviation:

(i) Standard deviation is used to determine the reliability of the means of two or more different series when these means are same.

(ii) It is widely used in other statistical measures like coefficient of skewness, correlation and regression analysis. It is very important statistical tool used in probability distribution.

1.2.Skewness and Kurtosis (Concept only)

Skewness:

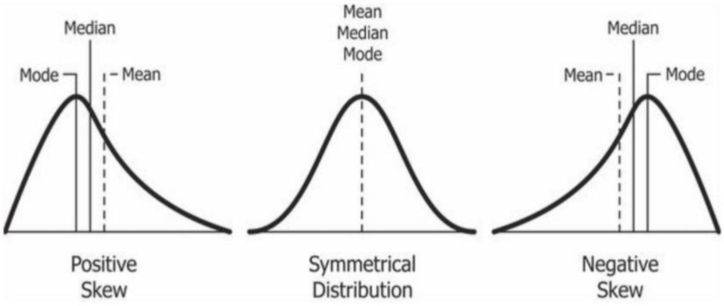
Literally skewness means 'lack of symmetry'. We study skewness to have an idea about the shape of the curve which we can draw with the help of given data. **A distribution is said to be skewed if (i) mean, median and mode fall at different points i.e, $\boldsymbol{mean \neq median \neq mode}$. (ii) quartiles are not equidistant from median and (iii) the curve drawn with the help of given data is stretched more to one side than to the other.**

Positive skewness and Negative skewness:

A distribution is said to have positive skewness, when mean > median > mode.

A distribution is said to have negative skewness, when mean < median < mode

Note: A distribution is said to be symmetric when
mean = median = mode



Measures of skewness:

(1) Karl Pearson's coefficient of skewness

$$Sk = \frac{Mean - Mode}{Standard\ deviation},$$ if the mode is not ill defined

$$Sk = \frac{3 \times (Mean - Median)}{Standard\ deviation},$$ if the mode is ill defined

(2) Bowley's Coefficient of skewness

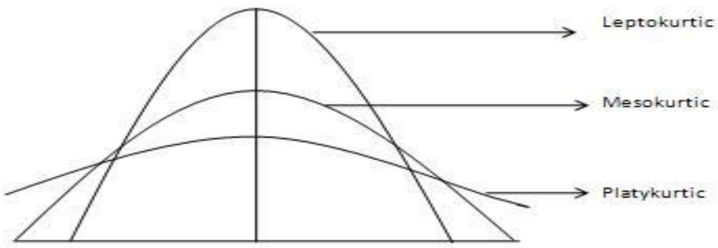
$$Sk = \frac{Q_3+Q_1-2\times median}{Q_3-Q_1}$$

Where, $Q_1 = first\ or\ lower\ quartile$

$Q_3 = third\ or\ higher\ quartile$

Kurtosis (Concept only)

Kurtosis means 'Convexity of a curve'. Kurtosis enables us to have an idea about the flatness or peakedness of the curve. The curve, which is neither peaked nor flat, is called **mesokurtic** or **normal curve** (as shown in the diagram). The curve, which is more peaked than the normal curve, is called **leptokurtic**. The curve, which is more flat than the normal curve is called **platykurtic**.



Ex.37. Given below the mean, the median and the standard deviation of two distributions. Determine which distribution is more skewed.

(i) Mean = 22, Median = 24, S.D. = 10

(ii) Mean = 22, Median = 25, S.D. = 12

Sol.

(i) Karl Pearson's coefficient of skewness is

$$Sk = \frac{3\times(Mean-Median)}{SD}$$

$$= \frac{3\times(22-24)}{10}$$

$$= -\frac{6}{10}$$

$$= -0.60$$

$$\therefore |Sk| = 0.60$$

(ii) Karl Pearson's coefficient of skewness is

$$Sk = \frac{3\times(Mean-Median)}{SD}$$

$$= \frac{3\times(22-25)}{12}$$

$$= -\frac{9}{12}$$

$$= -0.75$$

$$\therefore |Sk| = 0.75$$

Since, the absolute value of the coefficient of skewness of the distribution (ii) is greater than the absolute value of the coefficient of skewness of the distribution (i), so, the distribution (ii) is more skewed.

Ex.38. Of the distribution, Karl Pearson's coefficient of skewness is 0.32, standard deviation is 6.5 and mean is 29.6. Find median and mode of the distribution.

Sol.

We know that Karl Pearson's coefficient of skewness is $Sk = \frac{Mean-Mode}{SD}$

So, $0.32 = \frac{29.6-Mode}{6.5}$

- ⇨ 2.08 = 29.6 – Mode
- ⇨ Mode = 29.6 – 2.08
- ⇨ **Mode = 27.52**

By using the empirical relation,

$\boldsymbol{Mode = 3\times Median - 2\times Mean}$

- ⇨ 27.52 = 3 × $Median - 2 \times 29.6$
- ⇨ 27.52 = 3 × $Median - 59.2$
- ⇨ $3\times Median = 27.52 + 59.2$
- ⇨ Median = $\frac{86.72}{3}$

$= 28.91$

Check your progress

Ex.39. The mean, the mode and the Karl Pearson's coefficient of skewness of a

distribution are 120, 123 and -0.3 respectively. Find the standard deviation of the distribution. Ans: σ = 10]

Ex.40. In the distribution, mean = 65, median = 70 and coefficient of skewness = - 0.6, find (i) mode, (ii) coefficient of variation (CV). [Ans: mode = 80, CV = 38.46%]

Ex.41. From the data given below compute coefficient of skewness and comments on its value.

Profit (‘000₹) :	**10 – 12**	**12 – 14**	**14 – 16**	**16 – 18**	**18 – 20**	**20 – 22**	**22 – 24**
No. of companies:	**7**	**15**	**18**	**20**	**25**	**10**	**5**

Sol.

C. I	f	**Mid value** X	fX	fX^2
10 – 12	7	11	77	847
12 – 14	15	13	195	2535
14 – 16	18	15	270	4050
16 – 18	20 f_0	17	340	5780
L18 – 20	25 f_m	19	475	9025
20 – 22	10 f_1	21	210	4410
22 – 24	5	23	115	2645
Total	$\sum f$ = 100		$\sum fX$ =1682	$\sum fX^2$ =29292

Mean = $\frac{\sum fX}{\sum f}$

$= \frac{1682}{100} = 16.82$

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

$= 18 + \frac{25-20}{2\times 25-20-10} \times 2$

$= 18 + 0.5 = 18.5$

$SD(\sigma) = \sqrt{\frac{\sum fX^2}{\sum f} - \left(\frac{\sum fX}{\sum f}\right)^2}$

$= \sqrt{\frac{29292}{100} - \left(\frac{1682}{100}\right)^2}$

$= 3.16$

Therefore, the coefficient of skewness is

$$Sk = \frac{Mean-Mode}{SD}$$

$$= \frac{16.82-18.5}{3.16} = -0.532$$

Interpretation: Since, the value of the coefficient of skewness is negative, so, the distribution is negatively skewed.

Check your progress

Ex.42. Determine coefficient of skewness based on mean and median of the following distribution.

Income (₹)	**: 400 – 500**	**500 – 600**	**600 – 700**	**700 – 800**	**800 – 900**
Number of Employees	**: 8**	**16**	**20**	**17**	**3**

[Answer: - 0.112]

Ex.43. Calculate measure of skewness based on quartiles (Bowley's method) from the following data.

Variable	**: 1 – 5**	**6 – 10**	**11 – 15**	**16 – 20**	**21 – 25**	**26 – 30**	**31 – 35**
Frequency	**: 3**	**4**	**68**	**30**	**10**	**6**	**2**

[Answer: 0.25]

------xxxxx-----

UNIT 2: TIME SERIES ANALYSIS AND INDEX NUMBERS

Structure

2.1. Concept of Time Series

2.2. Applications of Time Series in Business decision making

2.3. Components of Time Series

2.4.Techniques of Time series analysis: Moving Average Method, Semi Average Method, and Least Square Method.

2.5. Meaning of Index Numbers

2.6. Constructions of price, quantity and value indices

2.7. Fixed base and Chain base indices

2.8. Uses of index numbers

2.1. Concept of Time Series

Statistical data arranged with respect to time are said to constitute a time series. For example, the series of values of price indices for the year 2001 to 2011 is a time series. Similarly, money deposited in a bank on various working days of a month is another example of a time series. **A time series represents the relationship between two variables of which the independent variable is time**. In the first example given above, different years starting from 2001 to 2011 are the values of the independent time variable. **Mathematically, the functional relationship between the two variables of a time series is expressed as $\boldsymbol{y = f(t)}$, y is the dependent variable which is a function of the independent variable time denoted by t. Clearly, a time series is a bivariate distribution where one variable is t and other is y, which is an observed value of an event at time t.** The values of t may be given in year, month, week, day, hour etc. Usually the values of t are given at equal intervals but sometimes the time intervals may be unequal.

Analysis of Time Series:

The methods of studying the effect of factors responsible for variations in time series data after separating them are collectively called analysis of time series.

The main objectives of time series analysis are:

(i) Identification of the causes of variation or fluctuation of time series data.

(ii) Separating the causes of variability, studying and analyzing each of them and finally estimating the effect of each of them on total variability of the series.

2.2. Applications of Time Series analysis in Business Decision Making

Time series analysis helps in analyzing the past, which comes in handy to forecast the future. The method is extensively employed in a financial and business forecast based on the historical pattern of data points collected over time and comparing it with the current trends. This is the biggest advantage used by organizations for decision making and policy planning by several organizations.

Time Series analysis is "an ordered sequence of values of a variable at equally spaced time intervals." It is used to understand the determining factors and structure behind the observed data, choose a model to forecast, thereby leading to better decision making. The Time Series Analysis is applied for various purposes, such as:

- Stock Market Analysis
- Economic Forecasting
- Inventory Studies
- Budgetary Analysis
- Census Analysis
- Yield Projection
- Sales Forecasting

2.3. Components of Time Series

Time series have four components which are described below:

(i) **Secular Trend (T):** Over a long period of time, if observed, most of the time series reveal either an inclining or a declining tendency. This general tendency of a time series over a fairly long period of time is termed as secular trend. It is also known as long term movement. Trend of a time series may be linear or non-linear.

The growth of population over a long period in the Country is an example of secular trend. The growth of production, profit of an organization, Literacy rate, mortality rate, death rate etc. are the examples of Secular trend.

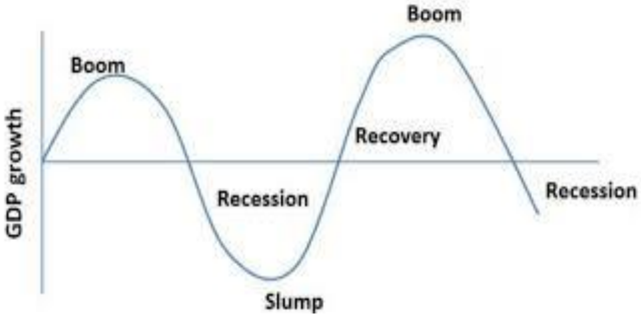
(ii) **Seasonal Variations (or Fluctuations) (S)**: The fluctuations or variations in the values of a time series exhibited over a period of one year or less are termed as seasonal fluctuations or seasonal variations. The seasonal fluctuations take place due to the following reasons:

(a) **Natural Causes**: Seasonal variations take place due to climatic changes. For example, during winter sale of wool increases, during summer demand for ice cream, cold drinks, electric fans, air cooler etc. increases; in rainy season demand for umbrella, rain coat etc. increases.

(b) **Rituals and Social Customs**: Man made rituals, social customs and traditions are also responsible for seasonal fluctuations of a time series. For example, just on the eve of New Year the sale of greeting cards, gift items, cake etc. increases to a great extent. In the beginning of the academic session sale of books, paper, uniforms etc. increase.

(iii) Cyclical variations (C): Most of the business and economic time series increase or decrease periodically with some amount of regularity. In general the periodicity of this type of variation is

more than one year. This periodic movement of a time series is termed as cyclic variations, for this happens due to business cycle. A business cycle has got four phases namely, prosperity or boom, recession, depression and recovery.



(iv) Random Variation or Irregular Movement (R or I): These variations are due to unpredictable factors which are sometimes even unidentifiable and which act more or less in a random manner. Some of such unpredictable factors include flood, famine, fire, epidemic, strike of trade unions terrorism etc. As these are unpredictable, they cannot be controlled by human hand.

2.4. Techniques of Time Series analysis

The following methods are used to measure secular trend

(i) Graphical Method (out of syllabus)

(ii) Moving Average Method

(iii) Semi-average Method

(iv) Least Squares Method (LSM)

Method of Semi Averages

In this method, we can find the solution of a secular trend. For this, we have to show our time series on graph paper. For example, we can take sales on X-axis and data of production on Y-axis. Now make the original graph by plotting points on graph paper with time and value pairs. After plotting original data we can calculate the trend line. For calculating the trend line, we will calculate semi-average. We divide the data into two equal parts with respect to time. And then we plot the arithmetic mean of the sets of values of Y against the center of the relative time span. If the number of observations is even then the division into halves will be done easily. But, for an odd number of observations, we will drop the middle most item, i.e$\frac{n+1}{2}$ th term. We need to join these two points together through a straight line which shows the trend. The trend values can then be read from the graph corresponding to each time period. Since extreme values greatly affect the arithmetic mean, and it is subjected to misleading values. Due to this, these trends may give distorted plots. But, if extreme values are not apparent, we may easily use and employ this method. To understand the estimation of trends, using the above mentioned two methods, consider the following working example.

Merits or advantages of Semi Average method

- This method is simple to understand as compare to other methods for measuring the secular trends.
- Everyone who applies this method will get the same result.

Demerits or disadvantages of Semi Average method

- The method assumes a straight line relationship between the plotted points without considering the fact whether that relationship exists or not.
- If we add more data to the original data then we have to do the complete process again for the new data to get the trend values and the trend line also changes.

Explanation of the Method

Here are two cases of calculating semi-average of data:

When data is even: In this case, the time series will be into two parts and then we calculate the average of each part. Suppose if we have 10 years data then we divide it into 5 -5 years and then we will calculate the first five-year average and the next five-year average after this we have to plot this on the graph paper. This will show the trend line as shown in the picture.

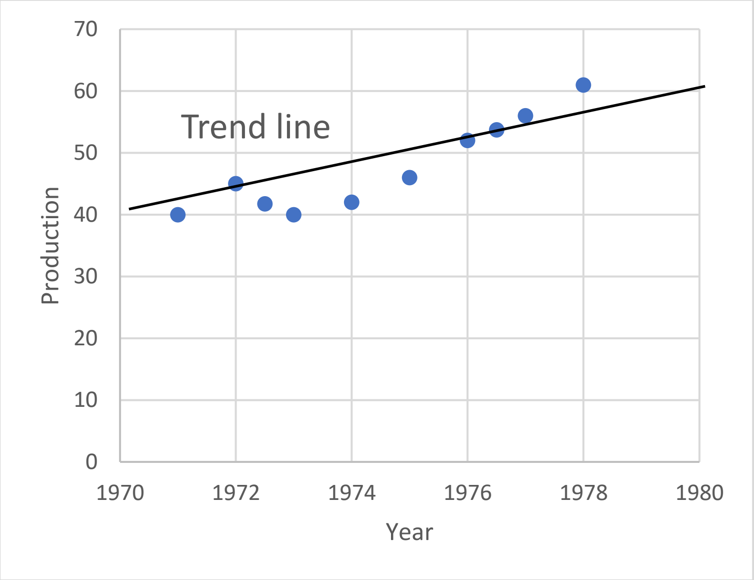
When data is odd: In this case, we just leave the middle data and we will follow the above-said procedure for the rest.

Solved Example on Method of Semi Averages

Year	Production	Semi-average
1971	40	

1972	45	
1972.5		**(40+45+40+42)/4=41.75**
1973	40	
1974	42	
1975	46	
1976	52	
1976.5		**(46+52+56+61)/4=53.75**
1977	56	
1978	61	

Thus we get two points 41.75 and 53.75 which we shall plot corresponding to their middle years i.e. 1972.5 and 1976.5. By joining these points we will obtain the required trend line.



Moving Average Method:

In this method of trend determination short term movements are estimated by taking moving averages (M.A.) of the values of the given time series. To determine secular trend by this method a knowledge about the length of cycles of cyclical fluctuations present in the series is most essential. It has been proved mathematically that in case of linear trend, seasonal and cyclical fluctuations are completed eliminated from the time series if period of moving average is equal to the average length of cycles. In finding moving averages each time a value is omitted from above and a value is included from below serially.

For ‘k’ observations, the moving average should be placed against the (k+1)th period. Omitting the 1st observation, the second ‘k’ moving average should be placed against the (k+2)th period and is continued till the last ‘k’ moving average.

Merits or advantages of Moving Average method

(i) It is easy to understand and simple to calculate.

(ii) There is some inherent flexibility in this method. Addition of a new values only increase the grand value and this has no effect on previous calculations.

(iii)If the period of moving average is equal to the average period of cycles, then this method completely eliminates the effect of cyclical fluctuations.

Demerits or disadvantages of Moving Average method

(i) In this method moving averages are lost in the beginning and at the end of the series

(ii) It is not suitable method for prediction purpose.

Ex.1. Calculate (i) three yearly, (ii) five yearly and (iii) four yearly moving averages for the following data and estimate sales for the year 2019.

Year	**: 2010**	**2011**	**2012**	**2013**	**2014**	**2015**	**2016**	**2017**	**2018**
Sales	**: 42**	**50**	**52**	**49**	**53**	**55**	**51**	**57**	**60**

Sol.

(i)

Year	Sales (y)	3 yearly moving total	3 yearly moving average
2010	42	---------	---------
2011	50	---------	---------
2012	52	---------	---------
2013	49	42+50+52 = 144	144/3=48
2014	53	50+52+49 = 151	151/4=37.75
2015	55	52+49+53 = 154	154/5=30.8
2016	51	49+53+55 = 157	157/3 = 52.33
2017	57	53+55+51 = 159	159/3 = 53
2018	60	55+51+57 = 163	163/7 = 23.29
2019		51+57+60 = 168	168/3=56

The estimated sales for the year 2019 is 56 units

(ii)

Year	**Sales (y)**	**4 yearly moving total**	**4 yearly moving average**
2010	42	---------	---------
2011	50	---------	---------
2012	52	---------	---------
2013	49	---------	--------
2014	53	42+50+52+49=193	193/4=48.25
2015	55	50+52+49+53=204	204/4=51
2016	51	52+49+53+55=209	209/4=52.25
2017	57	49+53+55+51=205	205/4=51.25
2018	60	53+55+51+57=216	216/4= 54
2019		55+51+57+60 =223	223/4=55.75

The estimated sales for the year 2019 is 55.75 units

(iii)

Year	Sales (y)	5-yearly moving total	5-yearly moving btotal
2010	42	-----------	-------
2011	50	-----------	-------
2012	52	----------	--------
2013	49	----------	
2014	53	-------	
2015	55	42+50+52+49+53=246	246/5 = 49.2
2016	51	50+52+49+53+55=259	259/4 = 64.75
2017	57	52+49+53+55+51=260	260/5=52
2018	60	49+53+55+51+57=265	265/5 = 53
2019		53+55++51+57+60=276	276/5 = 53.2

The estimated sales for the year 2019 is 53.2 units

Least Squares Method

If the variable y denotes the dependent variable and x, the time variable which is independent, the equation of the type $\boldsymbol{y = a + bx}$ is called linear trend or trend line, where, a and b are two parameters. By using the principle of least square method, a and b are estimated by solving the following two normal equations:

$$\sum y = na + b\sum x$$

$$\sum xy = a\sum x + b\sum x^2$$

If the values of time t are in equal interval, then consider $x = t - mid\ value\ of\ t$

So, $\sum x = 0$. In such case, we get $\boldsymbol{a = \frac{\sum y}{n}\ and\ b = \frac{\sum xy}{\sum x^2}}$

Ex.2. Production in a sugar mill is given below. Fit a linear trend by the method of least squares and estimate the production for the year 2016

Year :	**2008**	**2009**	**2010**	**2011**	**2012**	**2013**	**2014**
Production ('000 quintals):	**40**	**45**	**46**	**42**	**47**	**49**	**46**

Sol.

Let, y be the production ('000 quintals) and x, the time variable. Then the linear trend is $\boldsymbol{y = a + bx}$, --------(1) where, a and b are the two parameters.

Year (t)	**Production (y)**	$\boldsymbol{x = t - 2011}$	$\boldsymbol{x^2}$	$\boldsymbol{xy}$
2008	40	- 3	9	-120
2009	45	- 2	4	-90
2010	46	- 1	1	-46
2011	42	0	0	0
2012	47	1	1	47
2013	49	2	4	98
2014	46	3	9	138
Total	$\boldsymbol{\sum y = 315}$	$\boldsymbol{\sum x = 0}$	$\boldsymbol{\sum x^2 = 2}$	$\boldsymbol{\sum xy = 27}$

Since, $\sum x = 0$, so $a = \frac{\sum y}{n} = \frac{315}{7} = 45$

and $b = \frac{\sum xy}{\sum x^2} = \frac{27}{28} = 0.96$

Putting the value of a and b in the equation (1), we get the required estimated linear trend as

$\boldsymbol{y = 45 + 0.96x}$

For the year t = 2016, x = 2016 – 2011 = 5. So, putting x = 5 in $y = 45 + 0.96x$, we get

$y = 45 + 0.96 \times 5 = 49.8$

Hence, the estimated production for the year 2016, is 49.8 thousand quintals.

Ex.3. From the data given below fit a straight line trend by the method of least squares:

Year	**: 2005**	**2006**	**2007**	**2008**	**2009**	**2010**	**2011**	**2012**
Sale (₹ in lakh)	**: 6.7**	**5.3**	**4.3**	**6.1**	**5.6**	**7.9**	**5.8**	**6.1**

Estimate sale for the year 2013.

Sol.

If x denotes the time variable and y denotes the sale (₹ in lakh), then the straight line trend is $\boldsymbol{y = a + bx}$, where, a and b are the parameters.

Year	y	$x = year - 2008.5$	xy	x^2
2005	6.7	- 3.5	- 23.45	12.25
2006	5.3	- 2.5	- 13.25	6.25
2007	4.3	- 1.5	- 6.45	2.25
2008	6.1	- 0.5	3.05	0.25
2009	5.6	0.5	2.80	0.25
2010	7.9	1.5	11.85	2.25
2011	5.8	2.5	14.5	6.25
2012	6.1	3.5	21.35	12.25
Total	$\sum y = 47.8$	$\sum x = 0$	$\sum xy = 10.4$	$\sum x^2 = 42$

Since, $\sum x = 0$, so, $a = \frac{\sum y}{n} = \frac{47.8}{8} = 5.975$

and $b = \frac{\sum xy}{\sum x^2} = \frac{10.4}{42} = 0.248$

Hence, the required straight line trend is

$\boldsymbol{y = 5.975 + 0.248x}$

Now, for the year 2013, $x = 2013 - 2008.5 = 4.5$

Putting x = 4.5 in $y = 5.975 + 0.248 \times 4.5 = 7.091 \cong 7.1$

Hence, the estimated sale for the year 2013 is ₹7.1 lakhs

Merits / advantages of Least Square Method:

(i) Being a mathematical method, it is free from subjective error.

(ii) This method gives trend values for all the years.

(iii) This is the suitable method used for forecasting or prediction.

Demerits of Least Square Method:

(i) In comparison to other methods of trend determination this method is relatively complicated and time consuming.

(ii) Inclusion of a new value requires all calculations to be afresh which is not the case with moving average method.

(iii) Sometimes determining whether a linear or quadratic or exponential trend is suitable for a given time series really creates a problem and needs in depth knowledge for taking a decision.

Check your progress:

Ex.4. From the data given below fit a straight line trend by the method of least squares:

Year :	**2010**	**2011**	**2012**	**2013**	**2014**	**2015**	**2016**	**2017**
Production (thousand tones):	**12**	**15**	**22**	**26**	**32**	**41**	**39**	**45**

Estimate the production in 2018.

2.5. Meaning of Index Number

Some fundamental definitions of index numbers given by some well known statisticians at different times given below:

"Index numbers are devices for measuring differences in the magnitude of a group of related variables" – **Croxton and Cowden**

"Index number is a special type of average which provides a measurement of relative changes from time to time or from place to place" – **Wesell and Willet**

"Index number is a single ratio (usually in percentage) which measures the combined (i.e., averaged) change of several variables between different times, places or situations" – **A. M. Tuttle.**

"An index number is a statistical measure designed to show changes in a variable or a group of variables with respect to time, geographical location or other characteristics" – **Spiegel.**

By studying the definitions mentioned above, we may define "**An index number is a statistical device which is designed to measure the relative changes expressed in percentage of a group of variables with respect to time, geographical location or other characteristics"**. We compare the variables in the **current or given period** with the same in some past periods called the **base periods.**

Characteristics of Index Numbers:

(i) Index Numbers are specialized averages.

(ii) Index numbers study the effects of such factors which cannot be measured directly.

(iii) Index numbers bring out the common characteristics of a group items.

(iv) The changes measured with the help of index number can be either in relation to time or in relation to place.

Uses or advantages or importance of Index Numbers:

(i) Index numbers help in studying trends and tendencies of a series.

(ii) Index numbers help in policy formulation.

(iii) Index numbers help in measuring the purchasing power of money.

(iv) Index numbers help in deflating the various values.

(v) Index numbers act as an economic barometers.

Types of Index Numbers

(i) Price Index Numbers: Theses index numbers measure the general changes in the prices. They are further sub-divided into following classes:

(a) **Wholesale Price Index Numbers**: These index numbers reflect the changes as the general price level of a country.

(b) **Retail Price Index Numbers**: These indices reflect the general changes in the retail prices of various commodities such as consumption goods,

stocks and shares, etc. **Cost of Living Index Number (CLIN)** or **Consumer Price Index Number (CPIN)** is a specialized type of retail price index.

Quantity Index Numbers: These index numbers the changes in the volume of goods produced (manufactured), consumed or distributed, like indices of agricultural production, industrial production, exports, imports etc. With the help of Quantity Index Numbers, we can study the level of physical output in an economy.

(ii) Value Index Numbers: With these index numbers, we can study the change in total value (price multiplied by quantity) of production. The value index is denoted by V_{01} and it is given by $V_{01} = \frac{\sum p_1 q_1}{\sum p_0 q_0}$

Problems in the construction of an Index Number

(i) **Understanding of the purpose**: A clear understanding of the purpose for which index numbers are to be constructed is very essential.

(ii) **Selection of commodities**: The commodities selected should be fairly representative of the phenomenon under investigation. It should remain uniform in quantity from year to year.

(iii) **Selection of sources of data**: A decision has to be taken about the arrangement to be made for obtaining the price quotations of the commodities selected from various centers. In case of large population of markets, we should select the sample of the markets.

(iv) **Selection of base period**: The base period should be selected in such a way that the base period should be normal period, it should not be too distant from the given or current period, the index number of a base period is always 100. A base period may be fixed or changing.

(v) **Choice of an average**: Usually, arithmetic mean or geometric mean is used for constructing index number. Theoretically, geometric mean is the best average in the construction of an ideal index number.

(vi) **Selection of appropriate weights**: Since, all items are not equally importance and hence it is necessary to assign appropriate weights to show the varying importance of the different items. In case of price index number, quantity is considered as weight and in case of quantity index number, price is considered as weight.

(vii) **Selection of an appropriate formula**: There are a large number of formulae which can be used for constructing index numbers. The selection of an appropriate formula depends on the purpose of the index number and the data available.

Methods of constructing Index Numbers

1. Unweighted Index Numbers:

(i) **Simple Aggregate Method**

Let, p_1 = price per unit of an item in the current period

p_0 = price per unit of the item in the base period

q_1 = quantity of the same item in the current period

q_0 = quantity of the same item in the base period

In this method the price index number is given by $\boldsymbol{P_{01}} = \frac{\Sigma p_1}{\Sigma p_0} \times \mathbf{100}$

That is

$$\boldsymbol{P_{01}} = \frac{\textit{Total price of items in the current period}}{\textit{Total price of the items in the base period}} \times \mathbf{100}$$

Similarly, the quantity index number is

$$\boldsymbol{Q_{01}} = \frac{\Sigma q_1}{\Sigma q_0} \times \mathbf{100}$$

That is,

$$\boldsymbol{Q_{01}} = \frac{\textit{Total quantity of items in the current period}}{\textit{Total quantity of the items in the base period}} \times \mathbf{100}$$

(ii) **Simple Average Relative Method**

We define $\boldsymbol{P} = \frac{p_1}{p_0} \times \mathbf{100}$, where, P is called **price relative**. In this method, price index number is given by $\boldsymbol{P_{01}} = \frac{\Sigma P}{N}$, where, N = total number of items.

Similarly, quantity index number is

$\boldsymbol{Q_{01}} = \frac{\Sigma Q}{N}$, where, $\boldsymbol{Q} = \frac{q_1}{q_0} \times \mathbf{100}$ is called quantity relative.

This method, though simple, is not reliable and has the following limitations.

(i) The prices of various commodities may be quoted in different units.

(ii) The various commodities are weighted according to the magnitudes of their prices and accordingly commodities which are highly priced exert a greater influence on the value of the index than the commodities which are low-priced.

(iii) The relative importance of the various commodities is not taken into consideration.

2. Weighted Index Numbers

(i) Laspeyre's Index Numbers:

Laspeyre's price index number is given by

$$\boldsymbol{P_{01}^{La}} = \frac{\boldsymbol{\Sigma p_1 q_0}}{\boldsymbol{\Sigma p_0 q_0}} \times \mathbf{100}$$

Similarly, Laspyre's quantity index

$$\boldsymbol{Q_{01}^{La}} = \frac{\boldsymbol{\Sigma q_1 p_0}}{\boldsymbol{\Sigma q_0 p_o}} \times \mathbf{100}$$

This method is upward bias.

(ii) Paasche's Index Numbers:

Paasche's price index number is given by $\boldsymbol{P_{01}^{Pa}} = \frac{\boldsymbol{\Sigma p_1 q_1}}{\boldsymbol{\Sigma p_0 q_1}} \times \mathbf{100}$`

Similarly, Paasche's quantity index number is $\boldsymbol{Q_{01}^{Pa}} = \frac{\boldsymbol{\Sigma q_1 p_1}}{\boldsymbol{\Sigma q_0 p_1}} \times \mathbf{100}$

This method is downward bias.

(iii) Fisher's Ideal Index Numbers: The geometric mean (G.M.) of Laspeyre's and Paasche's index numbers, is called Fisher's Ideal Index Number.

Thus, Fisher's ideal price index number is given by $P_{01}^{F} = \sqrt{P_{01}^{La} \times P_{01}^{Pa}}$

Or $P_{01}^{F} = \sqrt{\frac{\boldsymbol{\Sigma p_1 q_0}}{\boldsymbol{\Sigma p_0 q_0}} \times \frac{\boldsymbol{\Sigma p_1 q_1}}{\boldsymbol{\Sigma p_0 q_1}}} \times 100$

Similarly, Fisher's ideal quantity index number is $\boldsymbol{Q_{01}^{F}} = \sqrt{\boldsymbol{Q_{01}^{La}} \times \boldsymbol{Q_{01}^{Pa}}}$

Or, $\boldsymbol{Q_{01}^{F}} = \sqrt{\frac{\boldsymbol{\Sigma q_1 p_0}}{\boldsymbol{\Sigma q_0 p_o}} \times \frac{\boldsymbol{\Sigma q_1 p_1}}{\boldsymbol{\Sigma q_0 p_1}}} \times \mathbf{100}$

(iv) Dorbish-Bowley Index Numbers: The arithmetic mean of Laspeyre's and paasche's index numbers is the Dorbish-Bowley's index number.

Thus, Dorbish-Bowley price index number is given by

$P_{01}^{DB} = \frac{1}{2} \times \left(\frac{\boldsymbol{\Sigma p_1 q_0}}{\boldsymbol{\Sigma p_0 q_0}} + \frac{\boldsymbol{\Sigma q_1 p_0}}{\boldsymbol{\Sigma q_0 p_o}}\right) \times \mathbf{100}$

Similarly, Dorbish-Bowley quantity index number is

$Q_{01}^{DB} = \frac{1}{2} \times \left(\frac{\boldsymbol{\Sigma q_1 p_0}}{\boldsymbol{\Sigma q_0 p_o}} + \frac{\boldsymbol{\Sigma q_1 p_1}}{\boldsymbol{\Sigma q_0 p_1}}\right) \times \mathbf{100}$

Ex.5. The following are the prices and quantities of commodities in two given in years 2005 and 2010. Calculate price index number for the year 2010:

(a) Laspeyre's method

(b) Paasche's method

(c) Dorbish-Bowley's method

(d) Fisher's ideal index.

	2005		**2010**	
Commdity	**Price**	**Quantity**	**Price**	**Quantity**
A	**4**	**50**	**10**	**40**
B	**3**	**10**	**9**	**2**
C	**2**	**5**	**4**	**2**

Sol.

2005		**2010**					
p_0	q_0	p_1	q_1	p_0q_0	p_1q_0	p_0q_1	p_1q_1
4	50	10	40	200	500	160	400
3	10	9	2	30	90	6	18
2	5	4	2	10	20	4	8
				$\sum p_0q_0$	$\sum p_1q_0$	$\sum p_0q_1$	$\sum p_1q_1$
				= 240	**= 610**	**= 170**	**= 426**

(a) Laspeyre's price index number

$$\mathrm{P}_{01}^{\mathrm{La}} = \frac{\sum \mathrm{p}_1\mathrm{q}_0}{\sum \mathrm{p}_0\mathrm{q}_0} \times 100 = \frac{610}{240} \times 100 = 254.17$$

(b) Paasche's price index number

$$P_{01}^{Pa} = \frac{\sum p_1q_1}{\sum p_0q_1} \times 100 = \frac{426}{170} \times 100 = 250.59$$

(c) Dorbish-Bowley's price index number

$$P_{01}^{DB} = \frac{1}{2} \times \left(\frac{\sum p_1q_0}{\sum p_0q_0} + \frac{\sum q_1p_0}{\sum q_0p_o}\right) \times \mathbf{100}$$

$$= \frac{1}{2} \times \left(\frac{610}{240} + \frac{426}{170}\right) \times 100$$

$$= \frac{1}{2} \times 504.754902$$

$$= 252.377451$$

(d) Fisher's ideal index number is

$$P_{01}^{F} = \sqrt{P_{01}^{La} \times P_{01}^{Pa}} = \sqrt{254.17 \times 250.59} = 252.374$$

Ex.6. From the data given below construct index number of the group of four commodities by using Fisher's Ideal formula:

	Base year		**Current year**	
Commo-dities	**Price per unit (₹)**	**Expenditure(₹)**	**Price per unit (₹)**	**Expenditure (₹)**
A	**2**	**40**	**5**	**75**
B	**4**	**16**	**8**	**40**
C	**1**	**10**	**2**	**24**
D	**5**	**25**	**10**	**60**

Sol.

p_0	p_0q_0	q_0	p_1	p_1q_1	q_1	p_0q_1	p_1q_0
2	40	$\frac{40}{2} = 20$	5	75	$\frac{75}{5} = 15$	30	100
4	16	$\frac{16}{4} = 4$	8	40	$\frac{40}{8} = 5$	20	32
1	10	$\frac{10}{1} = 10$	2	24	$\frac{24}{2} = 12$	12	20
5	25	$\frac{25}{5} = 5$	10	60	$\frac{60}{10} = 6$	30	50
	Σp_0q_0 = 91			Σp_1q_1 = 199		Σp_0q_1 =92	Σp_1q_0 = 202

Fisher's ideal index number = $\sqrt{\frac{\Sigma p_1q_0}{\Sigma p_0q_0} \times \frac{\Sigma p_1q_1}{\Sigma p_0q_1}} \times 100$

$$= \sqrt{\frac{202}{91} \times \frac{199}{92}} \times 100$$

$$= 219.12$$

Tests of Adequacy or Consistency of Index Number

Time Reversal Test (TRT):

Let, I_{01} be an index number of the current period 1 with respect to the base period 0 and I_{10} , the index number of the base period 0 with respect to the current period 1. If the index number satisfies the relation $\boldsymbol{I_{01} \times I_{10} = 1}$, then it is said to follow Time Reversal Test (TRT).

It can be proved that Fisher's ideal index formula follows TRT.

Factor Reversal Test (FRT):

Let, P_{01} and Q_{01}be respective the price and quantity index numbers of the current period 1 with respect to the base period 0. If the index number satisfies the relation $\boldsymbol{P_{01} \times Q_{01} = V_{01}}$, where, V_{01} is the value index number and $\boldsymbol{V_{01} = \frac{\Sigma p_1 q_1}{\Sigma p_0 q_0}}$, then the index number is said to follow Time Reversal Test (FRT). It can also be proved that Fisher's ideal index number satisfies FRT.

Important question: **Why is Fisher's index said to be an Ideal index?**

Ans: Fisher's index is said to be an idal index due to the following reasons:

i) Fisher's index satisfies Time Reversal Test and Factor Reversal Test.

ii) Since, this index is the geometric mean of Laspeyre's and Paasche's indices, so, is free from downward bias and upward bias.

Ex.7. From the data given below construct index number of the group of four commodities by using Fisher's Ideal formula and show that it satisfies Time Reversal Test and Factor Reversal Test.

	Base year		**Current year**	
Commo-dities	**Price per unit (₹)**	**Quantity (Kg)**	**Price per unit (₹)**	**Quantity (Kg)**
A	**2**	**20**	**5**	**15**
B	**4**	**4**	**8**	**5**
C	**1**	**10**	**2**	**12**
D	**5**	**5**	**10**	**6**

Sol.

p_0	q_0	p_1	q_1	p_0q_0	p_1q_0	p_0q_1	p_1q_1
2	20	5	15	40	100	30	75

4	4	8	5	16	32	20	40
1	10	2	12	10	20	12	24
5	5	10	6	25	50	30	60
				$\boldsymbol{\Sigma p_0 q_0}$	$\boldsymbol{\Sigma p_1 q_0}$	$\boldsymbol{\Sigma p_0 q_1}$	$\boldsymbol{\Sigma p_1 q_1}$
				= 91	**= 202**	**= 92**	**= 199**

Fisher's ideal index number is $\sqrt{\frac{\boldsymbol{\Sigma p_1 q_0}}{\boldsymbol{\Sigma p_0 q_0}} \times \frac{\boldsymbol{\Sigma p_1 q_1}}{\boldsymbol{\Sigma p_0 q_1}}} \times 100$

$$= \sqrt{\frac{202}{91} \times \frac{199}{92}} \times 100$$

$$= 219.12$$

Ignoring the factor 100, we have

$P_{01}^F = \sqrt{\frac{\boldsymbol{\Sigma p_1 q_0}}{\boldsymbol{\Sigma p_0 q_0}} \times \frac{\boldsymbol{\Sigma p_1 q_1}}{\boldsymbol{\Sigma p_0 q_1}}} = \sqrt{\frac{202}{91} \times \frac{199}{92}}$----------(1)

$P_{10}^F = \sqrt{\frac{\Sigma p_0 q_1}{\Sigma p_1 q_1} \times \frac{\Sigma p_0 q_0}{\Sigma p_1 q_0}} = \sqrt{\frac{92}{199} \times \frac{91}{202}}$ -----------(2)

Multiplying (1) and (2), we get

$$P_{01}^F \times P_{10}^F = \sqrt{\frac{202}{91} \times \frac{199}{92} \times \frac{92}{199} \times \frac{91}{202}} = \sqrt{1} = 1$$

Which shows that Fisher's index satisfies TRT

Ignoring the factor 100

$Q_{01}^F = \sqrt{\frac{\boldsymbol{\Sigma q_1 p_0}}{\boldsymbol{\Sigma q_0 p_0}} \times \frac{\boldsymbol{\Sigma q_1 p_1}}{\boldsymbol{\Sigma q_0 p_1}}} = \sqrt{\frac{92}{91} \times \frac{199}{202}}$ ---------- (3)

Multiplying (1) and (3), we get

$$P_{01}^F \times Q_{01}^F = \sqrt{\frac{202}{91} \times \frac{199}{92} \times \frac{92}{91} \times \frac{199}{202}}$$

$$= \sqrt{\frac{199^2}{91^2}}$$

$$= \frac{199}{91}$$

$$= \frac{\Sigma p_1 q_1}{\Sigma p_0 q_0}$$

$$= V_{01}$$

Hence, Fisher’s index satisfies FRT.

Cost of Living Index Number (CLIN) Or Consumers Price Index Number (CPIN):

Cost of living index number (CLIN) is the index number which shows the relative changes of cost of various items consumed by different classes of people in two different periods to maintain the same standard of living.

Steps of Construction of Cost of Living Index Number (CLIN):

The following are the main steps in constructing CLIN:

(i) **Selection of consumer class and area to be covered**: Decision should be taken regarding the class of people for whom the index number is meant and geographical area to be covered.

(ii) **To conduct a family budget enquiry**: The enquiry should be conducted on a sample of households selected by the process of random sampling. The items of consumption can be classified under the following major heads:

(a) Food

(b) Clothing

(c) Fuel and lighting

(d) Housing

(e) Miscellaneous.

Information should be collected in respect of price and quantity of different goods and service.

(iii) **To collect the retail prices of items**: The retail prices of the items are to be collected from the localities where the class of people concerned reside. Price quotations should be obtained at least once in a week.

(iv) **Selection of base year**: The base should be a year of economic stability and it should not be too distant from the current year.

Generally cost of living index is constructed for each week. The average of the weekly indices is taken as the index for a month. The average of monthly indices gives the cost of living index for the whole year.

Methods of Constructing cost of Living Index:

(i) Aggregate Expenditure Method or Weighted Aggregate Method:

This is Laspeyre’s method of constructing index numbers. The various commodities are assigned weights on the basis of quantity consumed in the base year. Base year prices and quantities are to be obtained from records.

$$\boldsymbol{CLIN = \frac{\sum p_1 q_0}{\sum p_0 q_0} \times 100}$$

$$\boldsymbol{= \frac{Total\ expenditure\ in\ the\ current\ year}{Total\ expenditure\ in\ base\ year} \times 100}$$

(ii)Family Budget Method or Method of Weighted Relatives

In this method, the cost of living index is calculated with the following formula:

$$\boldsymbol{CLIN = \frac{\sum PW}{\sum W}}$$

Where, $P = \frac{p_1}{p_0}$ = Price relative and $W = p_0 q_0$ = weights

OR

$$\boldsymbol{CLIN = \frac{\sum IW}{\sum W}}$$

Where, I = group index and W = weight.

Ex.8. An enquiry into budgets of middle class families in a city gave the following information:

Expenses on	**:Food**	**Rent**	**Clothing**	**Fuel**	**Miscellaneous**
	35%	**15%**	**20%**	**10%**	**20%**
Price in ₹ (2007)	**: 150**	**50**	**100**	**20**	**60**
Price in ₹ (2008)	**: 174**	**60**	**125**	**25**	**90**

What changes in the cost of living figure of 2008 have taken place as compared in 2007?

Sol.

Expenses on	**Price in 2007** $(\boldsymbol{p_0})$	**Price in 2008** $(\boldsymbol{p_1})$	$\boldsymbol{P = \frac{p_1}{p_0} \times 100}$	**W**	**PW**
Food	150	174	$\frac{174}{150} \times 100 = 116$	35	4060
Rent	50	60	$\frac{60}{50} \times 100 = 120$	15	1800
Clothing	100	125	$\frac{125}{100} \times 100 = 125$	20	2500
Fuel	20	25	$\frac{25}{20} \times 100 = 125$	10	1250
Miscellaneous	60	90	$\frac{90}{60} \times 100 = 150$	20	3000
Total				$\boldsymbol{\sum W}$ **= 100**	$\boldsymbol{\sum PW}$ **= 1260**

Cost of Living Index Number (CLIN) $= \frac{\sum PW}{\sum W}$

$$= \frac{12610}{100}$$

$$= 126.10$$

As compared to 2007, the changes in the cost of living figure in 2008 is 126.1% i.e., 26.1% more.

Ex.9. The following table gives the index numbers for different groups together with their respective weight for 2010 (base year = 2005)

Group	**Group Index Number**	**Group Weight**
Food	**130**	**60**
Clothing	**280**	**5**
Lighting & Fuel	**190**	**7**
Rent	**300**	**9**
Miscellaneous	**210**	**19**

(i) Find out the overall cost of living index number for the year 2010

(ii) Suppose a person was earning ₹5000 per month in 2005, what should be his salary in 2010, if his

standard of living in the year is to be same in 2005?

Sol.

Group	**Group Index (I)**	**Group weight (W)**	**IW**
Food	130	60	7800
Clothing	280	5	1400
Lighting & Fuel	190	7	1330
Rent	300	9	2700
Miscellaneous	200	19	3800
Total		$\boldsymbol{\sum W = 100}$	$\boldsymbol{\sum IW = 17030}$

(i) CLIN$=\frac{\sum IW}{\sum W}=\frac{17030}{100}=170.30$

(ii) The salary in 2010 is $\frac{CLIN\times salary\ in\ 2005}{Index\ Number\ in\ 2005}$

$$=\frac{170.30\times 5000}{100} = ₹8515$$

Fixed Base Index Number and Chain Base Index Number

Fixed Base Index Number

In fixed base method, a particular year is generally chosen arbitrarily and the prices of the subsequent years are expressed as relatives of the price of the base year. Sometimes instead of choosing a single year as the base, a period of a few years is chosen and the average price of this period is taken as the base year's price. The year which is selected as a base should be a normal year, or in other words, the price level in this year should neither be abnormally low nor abnormally high. If an abnormal year is chosen as the base, the price relatives of the current year calculated on its basis would give misleading conclusions. For example, the year 2020 in which pandemic due to Conid-19 was at its peak, is chosen as a base year; thus the comparison of the price level of the subsequent years to the price of 2020 is bound to give misleading conclusions as the price level in 2020 was abnormally high. In order to remove the difficulty associated with the selection of a normal year, the average price of a few years is sometimes taken as the base price. The fixed base method is used by the government in the calculation of national index numbers. In fixed base index number,

$$\boldsymbol{Price\ relatives\ in\ current\ year} = \frac{\boldsymbol{Price\ in\ current\ year}}{\boldsymbol{Price\ in\ base\ year}} \times \mathbf{100}$$

Merits or advantages of Fixed Base Index Number

- Fixed base indices are simple to understand and simple to calculate.
- There is no problem of constructing fixed base index numbers if data in respect of one or more years between the base and current years are missing.

Disadvantages of Fixed Base Index Number:

- Fixed Base Index Numbers become more and more inaccurate as the distance between the base period and the current period increases.
- Weights cannot be adjusted frequently.

Chain Base Index Number

In this method, there is no fixed base period; the year immediately preceding the one for which the price index has to be calculated is assumed as the base year. Thus, for the year 2014 the base year would be 2013, for 2013 it would be 2012, for 2012 it would be 2011, and so on. In this way there is no fixed base and it keeps on changing. The index numbers, calculated after taking the preceding year as the base year, are called Link Indices or Link Relatives.

Merits or advantages of Chain Base Index Numbers

- With the help of chain base indices we are able to know the gradual changes of prices or quantities from period to period.
- Weights can be adjusted as frequently as possible.
- Index numbers calculated by the chain-base method are free to a greater extent from seasonal variations than those obtained from fixed base method.
- The chain base method is very useful in respect of economic and business data where comparisons are to be made with previous period and not with any distant past.

Demerits or disadvantages of Chain Base Index Numbers:

- Chain base index numbers are not easy to understand as fixed base index numbers.

- The calculations involved in the construction chain base indices are tedious and laborious.
- Chain base indices cannot be computed if data for one or more periods are missing.

------xxxxx-----

UNIT 3: CORRELATION AND REGRESSION

Structure

3.1. Meaning and the types of correlation

In a bivariate distribution, if the change of one variable may effect the change of other variable, then the two variables are said to have correlation. For example: price and demand, heights and weights of persons etc.

Types of Correlation

(i) Positive, negative and zero correlations

(ii) Simple, multiple and partial correlations

(iii)Linear and non-linear correlations

Positive, negative and zero correlations:

In a bivariate distribution, if one variable increases (or decreases), the other variable also increases (or decreases) i.e, both the variables are in the same direction, then the two variables are said to have **positive correlation**. For example: income and expenditure, heights and weights of a group of persons, sales of ice-cream / cold drinks and the day temperature etc.

In a bivariate distribution, if one variable increases (or decreases), the other variable decreases (or increases) i.e, both the variables are in the opposite direction, then the two variables are said to have **negative correlation**. For example: price and demand of a certain commodities, the number of workers and the time required to complete the job etc.

In a bivariate distribution, if the change of one variable does not effect the change of other variable, then the two variables are said to have **zero correlation**. It can be proved that two independent variables are **uncorrelated**. For example: height and intelligence, price of sugar and demand of rice.

Simple, partial and multiple correlations

If we study the correlation between two variables, it is called **simple correlation**. If we study the correlation between more than two variables, it may be either partial or multiple correlation. In **partial correlation**, we measure the correlation between two variables (one dependent and other independent) when all other variables involve are kept constant. For example: if we limit our correlation analysis of yields and rainfall to periods where a certain average temperature existed, it will be a case of partial correlation.

In **multiple correlation**, the correlation between three or more than three variables are studied simultaneously. For example: if we study the yield of wheat per acre and both the amount of rainfall and fertilizers used, it is a problem of multiple correlation.

Linear and non-linear correlation:

If the amount of change in one variable tends to be a constant ratio to the amount of change in the other variable, then the correlation said to be a **linear**.

For example:

X:	2	3	5	7
Y:	4	6	10	14

Correlation would be called **non-linear** or **curvilinear**, if the amount of change in one variable does not tend to be a constant ratio to the amount of change in the other variable.

Methods of determination of correlation between two variables:

3.2 Karl Pearson's coefficient of correlation

It is a mathematical method to obtain the correlation between two variables. Karl Pearson's coefficient of correlation between the two variables X and Y is

$$\boldsymbol{r_{X,Y} \ or \ r(X,Y) \ or \ r = \frac{Covariance(X,Y)}{\sigma_X \times \sigma_Y}}$$

Where, $Covariance(X,Y) = \frac{\sum(X-\bar{X})(Y-\bar{Y})}{n}$

$\sigma_X = standard\ deviation\ of\ X = \sqrt{\frac{\sum(X-\bar{X})^2}{n}}$,

$\sigma_Y = standard\ deviation\ of\ Y = \sqrt{\frac{\sum(Y-\bar{Y})^2}{n}}$

Formula of $r_{X,Y}$

If we put the covariance and standard deviations formulae in the expression mentioned above and simplify, we get the following expressions which are used as a formula of correlation coefficient between the two variables X and Y.

(i) $\boldsymbol{r} = \frac{\sum(\boldsymbol{X}-\bar{\boldsymbol{X}})(\boldsymbol{Y}-\bar{\boldsymbol{Y}})}{\sqrt{\sum(\boldsymbol{X}-\bar{\boldsymbol{X}})^2 \times \sum(\boldsymbol{Y}-\bar{\boldsymbol{Y}})^2}}$

(ii) $\boldsymbol{r} = \frac{\sum \boldsymbol{XY} - \frac{(\sum \boldsymbol{X}) \times (\sum \boldsymbol{Y})}{\boldsymbol{n}}}{\sqrt{\left\{\sum \boldsymbol{X}^2 - \frac{(\sum \boldsymbol{X})^2}{\boldsymbol{n}}\right\} \times \left\{\sum \boldsymbol{Y}^2 - \frac{(\sum \boldsymbol{Y})^2}{\boldsymbol{n}}\right\}}}$

(iii) $\boldsymbol{r} = \frac{\sum \boldsymbol{UV} - \frac{(\sum \boldsymbol{U}) \times (\sum \boldsymbol{V})}{\boldsymbol{n}}}{\sqrt{\left\{\sum \boldsymbol{U}^2 - \frac{(\sum \boldsymbol{U})^2}{\boldsymbol{n}}\right\} \times \left\{\sum \boldsymbol{V}^2 - \frac{(\sum \boldsymbol{V})^2}{\boldsymbol{n}}\right\}}}$

where, $U = X - a, V = Y - b,$

$a = assumed\ mean\ of\ X$

$and\ b = assumed\ mean\ of\ Y$

Properties of Correlation Coefficient

(i) The correlation coefficient between two variables say X and Y always lies between -1 and +1 i.e, $-1 \leq r(X,Y) \leq +1$

(ii) The correlation coefficient between two variables are symmetric i.e, $r(X,Y) = r(Y,X)$

(iii) The correlation coefficient between two variables say X and Y is independent of the change of origin and scale. Thus, $if\ U = \frac{X-a}{h}$, $V = \frac{Y-b}{k}$, where, a, b, h, k ≠0, then $r(X,Y) = r(U,V)$

(iv) If two variables are independent, the correlation between the two variables is zero. In other words, two independent variables are uncorrelated. But the reverse is not always true.

Interpretations of the various values of r:

Since, the correlation coefficient between the two variables X and Y always lies from -1 to +1. That is, $-1 \leq r(X,Y) \leq 1$, so, the following are the interpretations of the various values of r.

(i) If r(X,Y) = -1, there is a perfect negative correlation between X and Y.

(ii) If – 1 < r(X,Y) < 0, there is a negative correlation between X and Y. In this case, the values of X and Y are in the opposite direction. For example: price and demand of a certain commodities.

(iii)If r(X,Y) = 0, X and Y are uncorrelated. For example: height and intelligence of a group of persons.

(iv)If 0 < r(X,Y) < 1, there is a positive correlation between X and Y. For example: demand and supply of a certain commodities.

(v) If r(X,Y) = 1, there is a perfect positive correlation between X and Y.

Ex.1. Find the correlation coefficient between the heights of the father and sons from the following data and interpret.

Heights of father (in inches) : 60 65 66 63 67 69 70

Heights of sons (in inches) : 65 64 66 62 69 68 69

Sol.

Let X and Y denotes the heights of father and son in inches respectively.

X	**Y**	**XY**	**X^2**	**Y^2**
60	65	3900	3600	4225
65	64	4160	4225	4096
66	66	4356	4356	4356
63	62	3906	3969	3844
67	69	4623	4489	4761
69	68	4692	4761	4624
70	69	4830	4900	4761
$\sum X$ **= 460**	**$\sum Y$** **= 463**	**$\sum XY$** **= 30467**	**$\sum X^2$** **= 30300**	**$\sum Y^2$** **= 30667**

Here, the number of pairs of the values (X,Y) is n = 7

The correlation coefficient between X and Y is

$$r(X,Y) = \frac{\sum XY - \frac{\sum X \times \sum Y}{n}}{\sqrt{\{\sum X^2 - \frac{(\sum X)^2}{n}\} \times \{\sum Y^2 - \frac{(\sum Y)^2}{n}\}}}$$

$$= \frac{30467 - \frac{460 \times 463}{7}}{\sqrt{\left\{30300 - \frac{460^2}{7}\right\} \times \left\{30667 - \frac{463^2}{7}\right\}}}$$

$$= \frac{41.2857}{\sqrt{71.4286 \times 42.8571}}$$

$$= \frac{41.2857}{55.3283} = 0.7462$$

Interpretation: There is a highly positive correlation between heights of fathers and sons.

Ex.2. Ten mechanics were asked to assemble a piece of machinery. The minutes X they took to assemble it in the morning and Y in the afternoon were observed and the following quantities were computed:

$\sum x = 142$, $\sum y = 166$, $\sum xy = 2434$, $\sum x^2 = 2085$, $\sum y^2 = 2897$

Determine the correlation coefficient between X and Y.

Sol.

The correlation coefficient between X and Y is

$$r = \frac{\sum XY - \frac{\sum X \times \sum Y}{n}}{\sqrt{\{\sum X^2 - \frac{(\sum X)^2}{n}\} \times \{\sum Y^2 - \frac{(\sum Y)^2}{n}\}}}$$

$$= \frac{2434 - \frac{142 \times 166}{10}}{\sqrt{\left\{2085 - \frac{142^2}{10}\right\} \times \{2897 - \frac{166^2}{10}\}}}$$

$$= \frac{76.8}{\sqrt{68.6 \times 141.4}} = 0.7798$$

[since, n = the number of mechanics = 10]

Ex.3. From the following data, compute coefficient of correlation between X and Y

X-series Y-series

Number of values	**15**	**15**
Mean	**25**	**18**
Sum of the square of the		
Deviations from the mean	**136**	**138**

Sum of the products of deviations of X and Y series from their respective means = 122

Sol.

Given, n = 15, $\bar{X} = 25, \bar{Y} = 18$

$\sum(X - \bar{X})^2 = 136, \sum(Y - \bar{Y})^2 = 138$ and

$\sum(X - \bar{X})(Y - \bar{Y}) = 122$

Therefore, the correlation coefficient between X and Y is

$$r = \frac{\sum(X-\bar{X})(Y-\bar{Y})}{\sqrt{\sum(X-\bar{X})^2 \times \sum(Y-\bar{Y})^2}}$$

$$= \frac{122}{\sqrt{136 \times 138}} = 0.8905$$

Ex.4. The coefficient of correlation between two variables X and Y is 0.48. The covariance is 36, the variance of X is 36. Find the standard deviation of Y.

Sol.

Given, r (X, Y) = 0.48, covariance (X, Y) = 36, $\sigma_X^2 = 36, \sigma_Y = ?$

By definition,

$$r(X,Y) = \frac{covariance(X,Y)}{\sigma_X \times \sigma_Y}$$

$$=> 0.48 = \frac{36}{\sqrt{36} \times \sigma_Y}$$

$$=> 0.48 = \frac{36}{6 \times \sigma_Y}$$

$$=> 0.48 = \frac{6}{\sigma_Y}$$

$$=> \sigma_Y = \frac{6}{0.48}$$

$$=> \sigma_Y = 12.5$$

Check your progress:

Ex.5. The correlation coefficient and covariance between X and Y is 0.25 and 3.6 respectively. If the variance of Y is 36, then find the standard deviation of X. (Ans: $\boldsymbol{\sigma_X = 2.4}$)

3.3 Spearman’s Rank Correlation

Karl Pearson’s coefficient of correlation discussed degree of linear relationship between two quantitative variables. But often we come across situation when definite measurements on the variables are not possible i.e., the factor, under study cannot be measured in quantitative terms. For example, the evaluation of a group of students on the basis of leadership ability, the ordering of women in beauty contest, the ranking of students by two or more judges in an interview and so on. In all such cases, objects are individuals may be ranked and arranged in order of merits or proficiency on two variables and the correlation between the ranks of two variables, is called **Rank correlation.**

In rank correlation, we may have three types of problems:

(i) When ranks are given

(ii) When ranks are not given

(iii)When equal ranks are given to more than two attributes.

(i) **When Ranks are given**

(a) Take the difference of two ranks D = $R_1 - R_2$

(b) Sum the square of D i.e., find $\sum D^2$

(c) Apply the formula $\boldsymbol{R = 1 - \frac{6\times\sum D^2}{n\times(n^2-1)}}$, where, n is the total number of pairs of observations.

(ii) When Ranks are not given

In this case, we assign the ranks of the values of the variables separately and apply the same process as in the case of when ranks are given.

(iii) **When equal ranks are given to more than two attributes**

In this case, we apply the formula,

$$\boldsymbol{R = 1 - \frac{6[\sum D^2 + \frac{1}{12}(m_1^3 - m_1) + \frac{1}{12}(m_2^3 - m_2) + \frac{1}{12}(m_3^3 - m_3) + \cdots]}{n(n^2-1)}}$$

Where, $m_1, m_2, m_3, \ldots$ *are the number of times a value is repeated.*

Ex.6. Two judges were asked to rank 7 different contestants. The ranks given by them

are given below.

Contestants:	**A**	**B**	**C**	**D**	**E**	**F**	**G**
Judge1 :	**2**	**1**	**4**	**3**	**5**	**7**	**6**
Judge2 :	**1**	**3**	**2**	**4**	**5**	**6**	**7**

Calculate Spearman's rank correlation coefficient.

Sol.

$\boldsymbol{R_1}$	$\boldsymbol{R_2}$	$\boldsymbol{D = R_1 - R_2}$	$\boldsymbol{D^2}$
2	1	1	1
1	3	-2	4
4	2	2	4
3	4	-1	1
5	5	0	0
7	6	1	1
6	7	-1	1
Total		**0**	$\boldsymbol{\sum D^2 = 12}$

Here,

n = 7, $\sum D^2 = 12$

Therefore,

$$R = 1 - \frac{6\times\sum D^2}{n\times(n^2-1)}$$

$$= 1 - \frac{6\times 12}{7\times(7^2-1)}$$

$$= 1 - 0.214 = 0.786$$

Ex.7. Find the Rank correlation coefficient between the heights of the father and sons from the following data.

Heights of father (in inches) :	**60**	**65**	**66**	**63**	**67**	**69**	**70**
Heights of sons (in inches) :	**65**	**64**	**66**	**62**	**69**	**68**	**67**

Sol.

Let, X be the height of father (in inches) and Y be the height of son (in inches)

X	Rank of X (R_1)	Y	Rank of Y (R_2)	D = R1 – R2	D^2
60	7	65	5	2	4

65	5	64	6	-1	1
66	4	66	4	0	0
63	6	62	7	-1	1
67	3	69	1	2	4
69	2	68	2	0	0
70	1	67	3	-2	4
				$\sum D = 0$	$\sum D^2 = 14$

Here,

n = 7, $\sum D^2 = 14$

Therefore,

$R = 1 - \frac{6 \times \sum D^2}{n \times (n^2 - 1)}$

$= 1 - \frac{6 \times 14}{7 \times (7^2 - 1)}$

$= 1 - 0.25 = 0.75$

Ex.8. From the following data of the marks obtained by 8 students in Statistics and Management paper, compute the rank coefficient of correlation

Marks in Statistics : 15 20 28 12 40 60 20 80

Marks in Management: 40 30 50 30 20 10 30 60

Sol.

Mathematics	R_1	**Statistics**	R_2	**D** **=R₁-R₂**	D^2
15	7	40	3	4	16
20 (5)	$\frac{5+6}{2} = 5.5$	30 (4)	$\frac{4+5+6}{3} = 5$	0.5	0.25
28	4	50	2	2	4
12	8	30 (5)	$\frac{4+5+6}{3} = 5$	3	9
40	3	20	7	-4	16
60	2	10	8	-6	36
20 (6)	$\frac{5+6}{2} = 5.5$	30 (6)	$\frac{4+5+6}{3} = 5$	0.5	0.25

80	1	60	1	0	0
Total				**0**	$\boldsymbol{\sum D^2}$ **=81.5**

We have, n = 8, m_1 = the number of repetition of 20 in Statistics = 2 and m_2 = the number of repetition of 30 in Management = 3

Therefore, Spearman's rank correlation is

$$R = 1 - \frac{6[\sum D^2 + \frac{1}{12}(m_1^3 - m_1) + \frac{1}{12}(m_2^3 - m_2)]}{n(n^2-1)}$$

$$= 1 - \frac{6\times[81.5 + \frac{1}{12}\times(2^3-2) + \frac{1}{12}\times(3^3-3)]}{8\times(8^2-1)}$$

$$= 1 - \frac{6\times 84}{8\times 63}$$

$$= 1 - 1 = 0$$

3.4 Meaning of Regression and Regression Analysis

Regression is a algebraic function of the average relationship between two kinds of variables. So, in a regression analysis, we concern two types of variables. The variable, which is predicted or estimated, is called **dependent variable**. The variable, which is used for prediction of the dependent variable, is called **independent variable**.

Regression Lines or Lines of Regression

Let us consider the two variables X and Y.

If X is **independent** and Y is **dependent** variables, then the equation of the type

$\boldsymbol{y = a_1 + b_1 x}$, is called the **regression line of Y on X**, where, $\boldsymbol{a_1}$ and $\boldsymbol{b_1}$ are the constants. $\boldsymbol{a_1}$is the **intercept** of Y and $\boldsymbol{b_1}$is the **slope** of the line. $\boldsymbol{b_1}$ is also known as the **regression coefficient of Y on X**.

If X is **dependent** and Y is **independent** variables, then the equation of the type

$\boldsymbol{x = a_2 + b_2 y}$ is called the **regression line of X on Y**, where, $\boldsymbol{a_2}$ is the **intercept of X** and $\boldsymbol{b_2}$ is the **slope of the line**. $\boldsymbol{b_2}$ is also known as the **regression coefficient of X on Y**.

Fitting a Regression Line

Fitting a regression line means to estimate the values of the parameters a and b.

Let us consider the regression line of Y on X. That is,

$\boldsymbol{y = a_1 + b_1 x}$

The value of the regression coefficient b_1 is obtained by using any one of the following formulae:

(i) $\boldsymbol{b_1} = \frac{\sum(\boldsymbol{x}-\bar{\boldsymbol{x}})(\boldsymbol{y}-\bar{\boldsymbol{y}})}{\sum(\boldsymbol{x}-\bar{\boldsymbol{x}})^2}$

(ii) $\boldsymbol{b_1} = \frac{\sum \boldsymbol{xy} - \frac{(\sum \boldsymbol{x})\times(\sum \boldsymbol{y})}{\boldsymbol{n}}}{\sum \boldsymbol{x}^2 - \frac{(\sum \boldsymbol{x})^2}{\boldsymbol{n}}}$

(iii) $\boldsymbol{b_1} = \boldsymbol{r} \times \frac{\boldsymbol{\sigma_y}}{\boldsymbol{\sigma_x}}$

Similarly, if we consider the regression line of X on Y i.e., $\boldsymbol{x = a_2 + b_2 y}$

The value of the regression coefficient b_1 is obtained by using any one of the following formulae

(i) $\boldsymbol{b_2} = \frac{\sum(\boldsymbol{x}-\bar{\boldsymbol{x}})(\boldsymbol{y}-\bar{\boldsymbol{y}})}{\sum(\boldsymbol{y}-\bar{\boldsymbol{y}})^2}$

(ii) $\boldsymbol{b_2} = \frac{\sum \boldsymbol{xy} - \frac{(\sum \boldsymbol{x})\times(\sum \boldsymbol{y})}{\boldsymbol{n}}}{\sum \boldsymbol{y}^2 - \frac{(\sum \boldsymbol{y})^2}{\boldsymbol{n}}}$

(iii) $\boldsymbol{b_2} = \boldsymbol{r} \frac{\boldsymbol{\sigma_x}}{\boldsymbol{\sigma_y}},$

The value of the intercept a_2 is obtained by the following formula

$\boldsymbol{a_2} = \bar{\boldsymbol{x}} - \boldsymbol{b_2} \times \bar{\boldsymbol{y}}$

Properties of Regression Coefficients

(i) $\sqrt{b_1 \times b_2} = r,$ the correlation coefficient between x and y

(ii) If $b_1(orb_2) \geq 1, the\ other\ must\ be\ b_2(orb_1) \leq 1$ so that $b_1 \times b_2 \leq 1$

(iii) The regression coefficients b_1 and b_2 are independent of the change of origin, but not on scale.

(iv) Both the regression coefficients b_1 and b_2 are positive or negative quantities. They cannot be one positive and other negative quantities.

(v) If both the regression coefficients are negative quantities, then the correlation coefficient must be negative quantity. If both the regression coefficients are positive quantities, the correlation coefficient must be positive quantity.

(vi) If the product of both the regression coefficients is 1 (or the correlation coefficient between x and y is 1), then both the regression lines are identical or coincide.

Ex.9. Given the bivariate data

X:	**2**	**4**	**5**	**6**	**8**	**11**
Y:	**18**	**12**	**10**	**8**	**7**	**5**

(i) **Fit a regression line of Y on X**

(ii) **Fit a regression line of X on Y**

(iii) **Predict Y, when X = 10**

(iv) **Predict X, when Y = 9**

(v) **Compute Karl Pearson's coefficient of correlation**

Sol.

Let the regression line of Y on X be $y = a_1 + b_1x$ and the regression line of X on Y be $x = a_2 + b_2y$

X	Y	XY	X^2	Y^2
2	18	36	4	324
4	12	48	16	144
5	10	50	25	100
6	8	48	36	64
8	7	56	64	49
11	5	55	121	25
$\sum X$= 36	$\sum Y$ = 60	$\sum XY$= 293	$\sum X^2$ =266	$\sum Y^2$ =706

(i) The regression coefficient of Y on X is

$$b_1 = \frac{\sum XY - \frac{(\sum X)\times(\sum Y)}{n}}{\sum X^2 - \frac{(\sum X)^2}{n}}$$

$$= \frac{293 - \frac{36\times60}{6}}{266 - \frac{36^2}{6}}$$

$= \frac{-67}{50} = -1.34$

$\bar{X} = \frac{\sum X}{n} = \frac{36}{6} = 6$ and $\bar{Y} = \frac{\sum Y}{n} = \frac{60}{6} = 10$

So, $a_1 = \bar{Y} - b_1\bar{X} = 10 - (-1.34) \times 6 = 18.04$

Therefore, the regression line of Y on X is

$\boldsymbol{Y = 18.04 - 1.34X}$

(ii) The regression coefficient of X on Y is

$$b_2 = \frac{\sum \boldsymbol{XY} - \frac{(\sum \boldsymbol{X}) \times (\sum \boldsymbol{Y})}{\boldsymbol{n}}}{\sum \boldsymbol{Y^2} - \frac{(\sum \boldsymbol{Y})^2}{\boldsymbol{n}}}$$

$$= \frac{293 - \frac{36 \times 60}{6}}{706 - \frac{60^2}{6}}$$

$$= -\frac{67}{106} = -0.632$$

So, $a_2 = \bar{X} - b_2\bar{Y}$

$= 6 - (-0.632) \times 10$

$= 12.32$

∴ The regression line of X on Y is $\boldsymbol{X = 12.32 - 0.632Y}$

(iii) To predict Y, we apply the regression line of Y on X. i.e., $\boldsymbol{Y = 18.04 - 1.34X}$

Putting X = 10, we get,

$Y = 18.04 - 1.34 \times 10 = 4.64$

Hence, the predicted value of Y is 4.64

(iv) To predict X, we apply the regression line of X on Y, i.e., $\boldsymbol{X = 12.32 - 0.632Y}$

Putting Y = 9, we get,

$X = 12.32 - 0.632 \times 9 = 6.632$

Hence, the predicted value of X is 6.632.

(v) By the property, we have,

$r = \sqrt{b_1 \times b_2} = \sqrt{(-1.34) \times (-0.632)}$

$= \pm 0.9203$

Since, both the regression coefficients are negative quantities, so, $r \neq 0.9203$

Hence, $r = -0.9203$

Ex.10. The following data are about sales and advertising expenditure of a firm:

	Sales **(₹ in crores)**	**Expenditure** **(₹ in crores)**
Mean	**40**	**6**
Standard deviation	**10**	**1.5**

Coefficient of correlation is $r = 0.9$

(i) Estimate the likely sales for a proposed advertisement expenditure of ₹ 10 crores.

(ii) What would be the advertisement expenditure if the firm fixes a sales target of 60 crores of rupees?

Sol.

Let X be the sales (₹ in crores) and Y be the advertising expenditure (₹ in crores)

(i) To estimate the likely sales i.e, X, we apply the regression line of X on Y. i.e.,

$X = a_2 + b_2Y$

Given, $\bar{X} = 40, \bar{Y} = 6, \sigma_X = 10, \sigma_Y = 1.5$

$and\ r = 0.9$

So, $b_2 = r \times \frac{\sigma_X}{\sigma_Y} = 0.9 \times \frac{10}{1.5} = 6$

$\therefore a_2 = \bar{X} - b_2\bar{Y} = 40 - 6 \times 6 = 4$

Therefore, the regression line of sales (X) on advertising expenditure (Y) is

$X = 4 + 6Y$

Putting Y = 10, we get

$X = 4 + 6 \times 10 = 64$

Hence, the estimated likely sales is ₹64 crores.

(ii) To estimate the advertisement expenditure i.e., Y, we apply the regression line of Y on X i.e.,

$Y = a_1 + b_1X$

$b_1 = r \times \frac{\sigma_Y}{\sigma_X} = 0.9 \times \frac{1.5}{10} = 0.135$

$\therefore a_1 = \bar{Y} - b_1\bar{X} = 6 - 0.135 \times 40 = 0.6$

Therefore, the regression line of advertisement expenditure (Y) on the sales (X) is
$Y = 0.6 + 0.135X$

Putting X = 60, we get
$Y = 0.6 + 0.135 \times 60 = 8.7$

Hence, the estimated advertisement expenditure is ₹8.7 crores.

Difference between Correlation And Regression

CORRELATION	REGRESSION
1. Correlation means the relationship between two variables	1. Regression is a mathematical measure expressing the average relationship between two or more variables.
2. Correlation coefficient between two variables are symmetric i.e, $r(X,Y) = r(Y,X)$	2. The regression coefficient between two variables are not symmetric i.e, $b_1 \neq b_2$
3. Correlation coefficient between two variables is independent of the change of origin and scale.	3. The regression coefficient between two variables is also independent of the change of origin, but not on scale.
4. The correlation coefficient between two variables always lies from -1 to +1 i.e, $-1 \leq r(X,Y) \leq 1$.	4. The values of the two regression coefficients must satisfy the relation $b_1 \times b_2 \leq 1$.

Why there are usually two lines of regression?

Let us consider the two variables X and Y. If X is an independent and Y is dependent variables then the equation of the type Y = a_1 + b_1X, is called the regression line of Y on X, where, a_1 and b_1 are two constants. This regression line is used to estimate the dependent variable Y, when the independent variable X is given. Again, if X is dependent and Y is independent variables, then the equation of the type X = a_2 + b_2Y, is called the regression line of X on Y, where, a$_2$ and b$_2$ are two constants. This regression line is used to estimate the dependent variable X, when the independent variable Y is given.

This is why there are usually two lines of regression.

Importance of Regression Analysis

(i) Regression analysis helps in establishing a functional relationship between two or more variables.

(ii) Since, most of the problems of Business and Economics analysis are based on cause and effect relationships, regression analysis is a highly valuable tool in economic and business research.

(iii) The regression analysis is very useful in prediction purposes.

3.5 Multiple Regression:

In multiple regression, we use more than one independent variables to estimate the dependent variable and as such multiple regression allows us to utilize more of the information available to us to estimate the dependent variable. Thus, **multiple regression is the average relationship between a dependent variable and two or more independent variables.**

Let us consider three variables X_1, X_2 and X_3. If X_1 be the dependent and X_2, X_3 be the independent variables, then **the regression line of X_1 on X_2 and X_3** is defined as

$\boldsymbol{X_1 = a_1 + b_{12.3}X_2 + b_{13.2}X_3}$, where, $a_1, b_{12.3}$ and $b_{13.2}$ are the parameters. These parameters are defined as follows and are obtained by applying the formulae given below

$\boldsymbol{b_{12.3}} =$ the partial regression coefficient of X_1 on X_2 when X_3 is kept constant $= \left(\frac{\mathbf{r_{12} - r_{13}r_{23}}}{\mathbf{1 - r_{23}^2}}\right)\frac{\mathbf{s_1}}{\mathbf{s_2}}$

$\boldsymbol{b_{13.2}} =$ the partial regression coefficient of X_1 on X_3 when X_2 is kept constant $= \left(\frac{\mathbf{r_{13} - r_{12}r_{32}}}{\mathbf{1 - r_{32}^2}}\right)\frac{\mathbf{s_1}}{\mathbf{s_3}}$

The value of the parameter a_1 is obtained by using the formula

$\boldsymbol{a_1 = \overline{X}_1 - b_{12.3}\overline{X}_2 - b_{13.2}\overline{X}_3}$

Similarly, If X_2 be the dependent variable and X_1, X_3 be the independent variables, then **the regression line of X_2 on X_1 and X_3** is given by

$\boldsymbol{X_2 = a_2 + b_{21.3}X_1 + b_{23.1}X_3}$**,** where,

$\boldsymbol{b_{21.3}}$ = the partial regression coefficient of X_2 on X_1 when X_3 is kept constant = $\left(\frac{\mathbf{r_{21}-r_{23}r_{13}}}{\mathbf{1-r_{13}^2}}\right)\frac{\mathbf{s_2}}{\mathbf{s_1}}$

$\boldsymbol{b_{23.1}}$ = the partial regression coefficient of X_2 on X_3 when X_1 is kept constant
$= \left(\frac{\mathbf{r_{23}-r_{21}r_{31}}}{\mathbf{1-r_{31}^2}}\right)\frac{\mathbf{s_2}}{\mathbf{s_3}}$

The value of the parameter a_2 is obtained by using the formula $\boldsymbol{a_2 = \bar{X}_2 - b_{21.3}\bar{X}_1 - b_{23.1}\bar{X}_3}$

If X_3 be the dependent variable and X_1, X_2 be the independent variables, then **the regression line of X_3 on X_1 and X_2** is

$\boldsymbol{X_3 = a_3 + b_{31.2}X_1 + b_{32.1}X_2,}$ where,

$\boldsymbol{b_{31.2}}$ = the partial regression coefficient of X_3 on X_1 when X_2 is kept constant = $\left(\frac{\boldsymbol{r_{31}-r_{32}r_{12}}}{\boldsymbol{1-r_{12}^2}}\right)\frac{\boldsymbol{s_3}}{\boldsymbol{s_1}}$

$\boldsymbol{b_{32.1}}$ = the partial regression coefficient of X_3 on X_2 when X_1 is kept constant = $\left(\frac{\boldsymbol{r_{32}-r_{31}r_{21}}}{\boldsymbol{1-r_{21}^2}}\right)\frac{\boldsymbol{s_3}}{\boldsymbol{s_2}}$

The value of the parameter a_3 is obtained by using the formula $\boldsymbol{a_3 = \bar{X}_3 - b_{31.2}\bar{X}_1 - b_{32.1}\bar{X}_2}$

Notations:

$$r_{12} = correlation\ coefficient\ between\ x_1 and\ x_2$$
$$r_{21} = correlation\ coefficient\ between\ x_2\ and\ x_1$$
$$r_{13} = correlation\ coefficient\ between\ x_1 and\ x_3$$
$$r_{31} = correlation\ coefficient\ between\ x_3\ and\ x_1$$
$$r_{23} = correlation\ coefficient\ between\ x_2 and\ x_3$$
$$r_{32} = correlation\ coefficient\ between\ x_3\ and\ x_2$$

By symmetric property,

$$r_{12} = r_{21}, r_{13} = r_{31}, r_{23} = r_{32}$$

$$s_1 = standard\ deviation\ of\ the\ variable\ x_1$$
$$s_2 = standard\ deviation\ of\ the\ variable\ x_2$$
$$s_3 = standard\ deviation\ of\ the\ variable\ x_3$$

Ex.11. For the following data, find the regression equation of X_3 on X_1 and X_2. Also estimate the value of X_3 when $X_1 = 4$ and $X_2 = 5$.

$$\boldsymbol{\bar{X}_1 = 6, \bar{X}_2 = 7, \bar{X}_3 = 8}$$
$$\boldsymbol{s_1 = 1, s_2 = 2, s_3 = 3}$$

$$\boldsymbol{r_{12} = 0.6, r_{13} = 0.7, r_{23} = 0.8}$$

Sol.

The regression equation of X_3 on X_1 and X_2 is given by

$X_3 = a_3 + b_{31.2}X_1 + b_{32.1}X_2$

Where,

$\mathrm{b}_{31.2} = \left(\frac{\boldsymbol{r_{31}-r_{32}r_{12}}}{\boldsymbol{1-r_{12}^2}}\right)\frac{\boldsymbol{s_3}}{\boldsymbol{s_1}} = \left(\frac{\mathbf{0.7-0.8\times0.6}}{\mathbf{1-0.6^2}}\right) \times \frac{\mathbf{3}}{\mathbf{1}} = 1.03$

$$\mathrm{b}_{32.1} = \left(\frac{\mathrm{r}_{32} - \mathrm{r}_{31}\mathrm{r}_{21}}{1 - \mathrm{r}_{21}^2}\right)\frac{\mathrm{s}_3}{\mathrm{s}_2} = \left(\frac{0.8 - 0.7 \times 0.6}{1 - 0.6^2}\right) \times \frac{3}{2} = 0.89$$

$\therefore a_3 = \bar{X}_3 - b_{31.2}\bar{X}_1 - b_{32.1}\bar{X}_2$

$= 8 - 1.03 \times 6 - 0.89 \times 7 = -4.41$

$\therefore$ The regression equation of X_3 on X_1 and X_2 is

$X_3 = -4.41 + 1.03X_1 + 0.89X_2$

Putting $X_1 = 4$ and $X_2 = 5$ we get

$X_3 = -4.41 + 1.03 \times 4 + 0.89 \times 5 = 4.16$

$\therefore$ The estimated value of X_3 is 4.16

Check your progress:

Ex.12. Find the regression equation of height of son on height of the father and estimate the height of a son when the height of a father is 68 inches.

Heights of father (in inches) :	**60**	**65**	**66**	**63**	**67**	**69**	**70**
Heights of sons (in inches) :	**65**	**64**	**66**	**62**	**69**	**68**	**69**

-----xxxxx-----

UNIT 4: PROBABILITY AND PROBABILITY DISTRIBUTION

Structure

Introduction: Concept of combination, Some important definitions

4.1. Definition of Probability and different approaches

4.2. Addition theorem

4.3. Multiplication theorem

4.4.Conditional probability and Baye's theorem

4.5. Random variable and probability distribution

4.6. Mathematical expectation and variance of a random variable

4.7. Probability distributions: Binomial distribution, Poisson distribution and Normal distribution

Introductions

Concept of Combination

Combination is the selection of 'n' objects by taking 'r' at a time. It is denoted by $\boldsymbol{n}_{\boldsymbol{C}_{\boldsymbol{r}}} = \frac{\boldsymbol{n}!}{\boldsymbol{r}!\times(\boldsymbol{n}-\boldsymbol{r})!}$

Where,

n! (factorial n) = $n \times (n-1) \times (n-2) \times \ldots \times 3 \times 2 \times 1$

r! (factorial r) = $r \times (r-1) \times (r-2) \times \ldots \times 3 \times 2 \times 1$

(n-r)! (factorial n-r)

= $(n-r) \times (n-r-1) \times (n-r-2) \times \ldots \times 3 \times 2 \times 1$

E,g., $2! = 2 \times 1 = 2$

$3! = 3 \times 2 \times 1 = 6$

$4! = 4 \times 3 \times 2 \times 1 = 24$

$5! = 5 \times 4 \times 3 \times 2 \times 1 = 120$

And so on…

Consider 4 objects say a, b, c and d

Select 4 objects by taking 2 at a time. In other words select 2 objects out of 4 objects. So, n = 4, r = 2. The possible selections are

ab, ac, ad

bc, bd, cd

The no. of selection is 6

By using the concept of combination, the number of selection is

$$4_{C_2} = \frac{4!}{2! \times (4-2)!} = \frac{4 \times 3 \times 2 \times 1}{2 \times 1 \times 2 \times 1} = 6$$

Some important results

(i) $\boldsymbol{n_{C_0} = n_{C_n} = 1}$ e.g $1_{C_0} = 1_{C_1} = 2_{C_0} = 2_{C_2} = 3_{C_0} = 3_{C_3} = \cdots = 1$

(ii) $\boldsymbol{n_{C_r} = n_{C_{n-r}}}$ **e.g** $1_{C_1} = 1_{C_0}, 3_{C_1} = 3_{C_2}, 4_{C_1} = 4_{C_3}, 4_{C_4} = 4_{C_0}, \ldots so\ on$

(iii)$\boldsymbol{n_{C_1} = n_{C_{n-1}} = n}$ e.g. $3_{C_1} = 3_{C_2} = 3, 5_{C_1} = 5_{C_4} = 5$ and so on

Shortcut method of the calculation of $\boldsymbol{n_{C_r}}$

If $\frac{n}{2} > r$,

$$\boldsymbol{n_{C_r} = \frac{n \times (n-1) \times (n-2) \times \ldots \times (n-r)}{r \times (r-1) \times (r-2) \times \ldots \times 3 \times 2 \times 1}}$$

If $\frac{n}{2} \leq r$,

$$\boldsymbol{n_{C_r} = n_{C_{n-r}}} = \frac{\boldsymbol{n \times (n-1) \times \ldots \times (n-r+1)}}{(n-r) \times (n-r-1) \times \ldots \times 3 \times 2 \times 1}$$

E.g.,

i) $8_{C_3} = \frac{8 \times 7 \times 6}{3 \times 2 \times 1} = 56$ $[\because \frac{8}{2} = 4 > 3]$

ii) $9_{C_6} = 9_{C_{9-6}} = 9_{C_3} = \frac{9 \times 8 \times 7}{3 \times 2 \times 1} = 84$ $[\because \frac{9}{2} = 4.5 < 6]$

iii) $10_{C_6} = 10_{C_{10-6}} = 10_{C_4} = \frac{10 \times 9 \times 8 \times 7}{4 \times 3 \times 2 \times 1} = 210$

$[\because \frac{10}{2} = 5 < 6]$

Some important definitions:

Random experiment: The experiment which has no unique result(s) but possibility of results are known in advance, is called a random experiment.

Tossing unbiased coin(s), throwing die (dice), drawing a card from a pack of 52 cards etc. are the examples of random experiments.

Sample space: If the possible occurrences of a random experiment are represented by a certain points, the collection or a set of all these points, is called sample space. It is denoted by S or Ω.

For example: Throwing an unbiased coin the sample space is S = {H, T}, throwing two unbiased coins the sample space is S = {HH, HT, TH, TT}, throwing three unbiased coins, the sample space is S = {HHH. HHT, HTH, HTT, THH, THT, TTH, TTT}, throwing a die, the sample space is S = {1, 2, 3, 4, 5, 6}, throwing two dice, the sample space is S = {(1,1), (1,2), (1,3), (1,4), (1,5), (1,6), (2,1), (2,2), (2, 3), (2,4), (2,5), (2,6), (3,1), (3,2), (3,3), (3,4), (3,5), (3,6), (4,1), (4,2), (4,3), (4,4), (4,5), (4,6), (5,1), (5,2), (5,3), (5,4), (5,5), (5,6), (6,1), (6,2), (6,3), (6,4), (6,5), (6,6)}

Event: Event means any result of a random experiment. Mathematically, an event is a sub set of a sample space.

For example: In tossing an unbiased coin {H} and {T} are the events. In throwing a die {1}, {2}, {3}, {4}, {5}, {6} are the events.

Exhaustive events: All the possible outcomes of a random experiment, is called the exhaustive events. In other words, the outcomes $A_1, A_2, \dots, A_k$ are called the exhaustive events if $A_1 \cup A_2 \cup A_3 \dots \cup A_k = S$, the sample space.

For example: In tossing 'n' unbiased coins, the number of exhaustive outcomes is 2^n

In throwing 'n' fair dice, the number of exhaustive events is 6^n. In drawing 'n' cards from a pack of 52 cards, the number of exhaustive events is 52_{C_n}

Equally likely events: Two or more events are said to be equally likely, if the chance of occurring each event is equal.

Mutually Exclusive events: Two or more events are said to be mutually likely, if any of them occurs, others do not occur. In other words, the events A and B are mutually exclusive, if $A \cap B = \phi$

4.1. Definition of Probability and different approaches

Probability has been defined in three different approaches which are given below.

Classical / Mathematical approach

Let, S be the sample space of a random experiment. Then n(S) is the number of all the possible outcomes. Let A be an event of the same random experiment. Then, n(A) is the number of favourable outcomes of the event A. The probability or chance of the event A, is defined as

$$P(A) = \frac{n(A)}{n(S)}$$

$$= \frac{No.of\ favourable\ outcomes\ of\ the\ event\ A}{Total\ no.of\ possible\ oucomes\ of\ the\ random\ experiment}$$

Limitations of Classical approach

If the events are not equally likely and the number of possible outcomes is not finite, we cannot apply the classical definition.

Empirical / Statistical / Frequency approach

Suppose a random experiment is made for n times out of which an event say A occurs r times, then the probability of the event a is given by $P(A) = \lim_{n \to \infty} \frac{r}{n}$.

Limitations of Empirical approach

The empirical definition of probability never be obtained in practice. The experimental conditions may not remain same when a random experiment is repeated a large number of times and the relative frequency $\frac{r}{n}$ may not remain unique limiting when $n \to \infty$

Axiomatic approach

To every event A, there corresponds a real valued function say P(A). The function P(A) is called probability of the event A, if it satisfies the following three axioms:

(i) $0 \leq P(A) \leq 1$

(ii) $P(\phi) = 0, P(S) = 1$

(iii) If $A_1, A_2, \dots, A_n$ be mutually exclusive events, then

$$P(A_1 \cup A_2 \dots \cup A_n) = P(A_1) + P(A_2) + \cdots + P(A_n)$$

Important results

(i) If ϕ is impossible event, the P(ϕ) = 0

(ii) If A^c be the complement of the event A, then $P(A^c)+P(A) = 1$ or $P(A^c) = 1 - P(A)$ or $P(A) = 1 - P(A^c)$

Ex.1. Two fair coins are tossed at random. What is the probability of getting (i) a head, (ii) two heads, (iii) at least one head, (iv) no head.

Sol.

If two fair coins are tossed, the sample space is

S = {HH, HT, TH, TT}

∴ The total number of outcomes n(S) = 4

(i) Let, A = Event of getting a head

= {HT, TH}

∴ The number of favourable outcomes n(A) = 2

The required probability is P(A) = $\frac{n(A)}{n(S)}$

= $\frac{2}{4}$

= $\frac{1}{2}$

= 0.5

(ii) Let, A = Event of getting two heads

= {HH}

∴ The number of favourable outcomes n(A) = 1

The required probability is P(A) = $\frac{n(A)}{n(S)}$

= $\frac{1}{4}$ = 0.25

(iii) Let, A = Event of getting at least one head, i.e., one or two heads

= {HH, HT, TH}

∴ The number of favourable outcomes n(A) = 3

The required probability is P(A) = $\frac{n(A)}{n(S)}$

$= \frac{3}{4} = 0.75$

(iv) Let, A = Event of getting no head

= {TT}

∴ The number of favourable outcomes n(A) = 1

The required probability is P(A) = $\frac{n(A)}{n(S)}$

$= \frac{1}{4} = 0.25$

Check your progress

Ex.2. Three fair coins are tossed at random. Find the probability of getting (i) a head, (ii) two heads, (iii) three heads, (iv) no head, (v) at least one head, (vi) at least two heads. [Answer: (i) $\frac{3}{8}$, (ii) $\frac{3}{8}$, (iii) $\frac{1}{8}$, (iv) $\frac{1}{8}$, (v) $\frac{7}{8}$, (vi) $\frac{1}{2}$

Ex.3. Two dice are thrown at random. What is the probability that (i) the sum of the numbers shown by the two dice is 9 ? (ii) the difference of the numbers shown by the two dice is 3?

Sol.

If two dice are thrown at random, the total number of possible outcomes is n(S) = $6^2 = 36$

(i) Let, A = Event that the sum of the numbers shown by the two dice is 9

={(3,6), (4,5), (5,4), (6,3)}

∴ The number of favourable outcomes is n(A) = 4

The required probability is

$$P(A) = \frac{n(A)}{n(S)} = \frac{4}{36} = \frac{1}{9}$$

(ii) Let, A = Event that the difference of the numbers shown by the two dice is 3

= {(1,4),(2,5),(3,6),(4,1),(5,2),(6,3)}

∴ The number of favourable outcomes is n(A) = 6

The required probability is

$$P(A) = \frac{n(A)}{n(S)} = \frac{6}{36} = \frac{1}{6}$$

Check your progress

Ex.4. Two dice are thrown at random. What is the probability that the numbers appeared on the uppermost faces of the two dice are same. [Answer:$\frac{1}{6}$]

4.2. Addition theorem

Case I: the events are not mutually exclusive.

The probability of occurring two events say A and B which are not mutually exclusive, is equal to the sum of the respective probabilities minus the probability of occurring two events A and B simultaneously. Thus,

$P(A \cup B) = P(A) + P(B) - P(A \cap B)$ or

P(A or B) = P(A) + P(B) – P(A and B)

If we consider for three events A, B and C, which are not mutually exclusive, then

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)$$

Case II: the events are mutually exclusive.

The probability of occurring 'k' mutually exclusive events say $A_1, A_2, \ldots, A_k$ is equal to the sum of their respective probabilities. Thus,

$$P(A_1 \cup A_2 \cup \ldots \cup A_k) = P(A_1) + P(A_2) + \cdots + P(A_k)$$

Ex.5. Out of 100 students in a hostel 80 take tea, 40 take coffee and 25 take both. Find the probability of a student (i) taking either tea or coffee (ii) not taking tea or coffee.

Sol.

Let, A = student taking tea, B = student taking coffee.

Therefore, $A \cap B =$ student taking both tea and coffee

So,

$P(A) = \frac{80}{100} = 0.8$

$P(B) = \frac{40}{100} = 0.4$

$P(A \cap B) = \frac{25}{100} = 0.25$

(i) P (taking either tea or coffee)

$= P(A \cup B) = P(A) + P(B) - P(A \cap B)$

$= 0.8 + 0.4 - 0.25$

$= 0.95$

(ii)P (not taking either tea or coffee)

$= P(A \cup B)^c$

$= 1 - P(A \cup B)$

$= 1 - 0.95$

$= 0.05$

Ex.6. From a set of 20 similar marbles marked from 1 to 20, a marble is drawn at random, find is the probability that its number multiple of (i) 2 or 5 (ii) 3 or 7.

Sol.

If a marble is drawn from the set of given 20 marbles, the total number of possible outcomes is $n(S) = 20$

(i) Let, A = the marble drawn is a multiple of 2

= $\{2, 4, 6, 8, 10, 12, 14, 16, 18, 20\}$

Let, B = the marble drawn is a multiple of 5

= {5, 10, 15, 20}

$\therefore A \cup B = \{2, 4, 5, 6, 8, 10, 12, 14, 15, 16, 18, 20\}$

$n(A \cup B) = 12$

P(multiple of 2 or 5) = P($A \cup B$)

$= \frac{n(A \cup B)}{n(S)}$

= $\frac{12}{20} = \frac{3}{5}$ = 0.6

(ii) Let, A = the marble drawn is a multiple of 3

= {3, 6, 9, 12, 15, 18}

Let, B = the marble drawn is a multiple of 7

= {7, 14}

$\therefore A \cup B = \{3, 6, 7, 9, 12, 15, 18\}$

$n(A \cup B) = 7$

P(multiple of 3 or 7) = P(A∪ B)

= $\frac{n(A \cup B)}{n(S)}$

$= \frac{7}{20}$

Ex.7. A man draws at random two balls from a bag containing 6 red and 5 white balls. Find the probability of getting the balls

(i) all are red

(ii) all are white

(iii) one is red and the other is white

(iv) are of same colours.

Sol.

Given, 6 red + 5 white = 11 balls

n(S) = the total number of outcomes that 2 balls are drawn out of 11 balls

$= 11_{C_2}$

$= \frac{11\times10}{2\times1}$

= 55

(i) Let, A = the balls drawn are red, i.e., two balls drawn are red.

∴ n(A) = the number of outcomes that 2 red balls are drawn from 6 red balls.

$= 6_{C_2}$

$= \frac{6\times5}{2\times1}$

= 15

∴ P(all are red) =P(A) = $\frac{n(A)}{n(S)}$

$= \frac{15}{55} = \frac{3}{11}$

(ii) Let, B = the balls drawn are white, i.e., two balls drawn are white.

∴ n(B) = the number of outcomes that 2 white balls are drawn from 5 white balls.

$= 5_{C_2}$

$= \frac{5\times4}{2\times1}$ = 10

∴ P(all are white) = P(B) = $\frac{n(B)}{n(S)}$

$= \frac{10}{55} = \frac{2}{11}$

(iii) Let, E = the balls drawn are one red and one white

∴ n(E) = the number of outcomes that one red ball is drawn from 6 red balls and one white ball is drawn from 5 white balls.

$= 6_{C_1} \times 5_{C_1}$

$= 6 \times 5 = 30$

∴ P(one is red and the other is white) = P(E)

$$= \frac{n(E)}{n(S)}$$

$$= \frac{30}{55} = \frac{6}{11}$$

(iv) P(the balls drawn are of same colours)

= P(the balls drawn are either 2 red or 2 white balls)

= P(A ∪ B)

= P(A) + P(B) – P($A \cap B$)

$= \frac{3}{11} + \frac{2}{11} - 0 = \frac{5}{11}$

4.3. Multiplication theorem

Case I: Events are not independent

For two events A and B which are not independent, the probability of occurring the events A and B, is equal to the the probability of occurring any one of them multiplied by the conditional probability of other event.

Thus,

$P(A \cap B)$ or P(A and B) or P(AB)$= P(A) \times P(B|A)$

$$= P(B) \times P(A|B)$$

Case II: Events are independent

The probability of occurring two events A and B is equal to the product of the respective probabilities. Thus,

$P(A \cap B)$ or P(A and B) or P(AB) $= P(A) \times P(B)$

Ex.8. A bag contains 6 white and 4 black marbles. Two marbles are drawn at random (a) without replacement (b) with replacement. Find the probability that

(i) the marbles drawn are white

(ii) the marbles drawn are black

(iii) the first marble drawn is white and the second is black.

Sol.

Given, 6 white + 4 black = 10 marbles

Two marbles are drawn at random successively.

(i) Let, A = the 1st marble drawn is a white

B = the 2nd marble drawn is also white

To find P(A∩B)

(a) In case of without replacement, A and B are not independent events.

So, the required probability is

P(A∩B)=P(A) × P(B|A)

$$= \frac{n(A)}{n(S_1)} \times \frac{n(B|A)}{n(S_2)}$$

$$= \frac{6_{C_1}}{10_{C_1}} \times \frac{5_{C_1}}{9_{C_1}}$$

$$= \frac{6}{10} \times \frac{5}{9} = \frac{1}{3}$$

Where, n(A) = the number of outcomes that the 1st white marble is drawn from 6 white marbles

$= 6_{C_1} = 6$

$n(S_1)$ = the number of outcomes that the 1st marble is drawn from 10 marbles.

$= 10_{C_1} = 10$

n(B|A) = the number of outcomes that the 2nd white marble is drawn from the remaining white marbles.

$= 5_{C_1} = 5$

$n(S_2)$ = the number of outcomes that the 2nd marble is drawn from the remaining 9 marbles.

$= 9_{C_1} = 9$

(b) In case of with replacement, A and B are independent events.

So, the required probability is

$P(A \cap B) = P(A) \times P(B)$

$$= \frac{n(A)}{n(S)} \times \frac{n(B)}{n(S)}$$

$$= \frac{6_{C_1}}{10_{C_1}} \times \frac{6_{c_1}}{10_{c_1}}$$

$$= \frac{6}{10} \times \frac{6}{10}$$

$$= \frac{9}{25}$$

Where,

n(A) = the number of outcomes that the 1st white marble is drawn from 6 white marbles

= 6_{C_1}= 6

n(B) = the number of outcomes that the 2nd white marble is also drawn from 6 white marbles

= 6_{C_1}= 6

n(S) = the number of outcomes that one (1st or 2nd) marble is drawn from 10 marbles.

= 10_{C_1}= 10

(ii) Let, A = the 1st marble drawn is black

B = the 2nd marble drawn is black

To find P(A∩ B)

(a) In case of without replacement, A and B are not independent events.

So, the required probability is

P(A∩B)=P(A) × P(B|A)

$$= \frac{n(A)}{n(S_1)} \times \frac{n(B|A)}{n(S_2)}$$

$$= \frac{4_{c_1}}{10_{c_1}} \times \frac{3_{c_1}}{9_{C_1}}$$

$$= \frac{4}{10} \times \frac{3}{9}$$

$$= \frac{2}{15}$$

where,

n(A) = the number of outcomes that the 1st black marble is drawn from 4 black marbles

= 4_{C_1}= 4

n(B|A) = the number of outcomes that the 2nd black marble is drawn from the remaining 3 black marbles.

= 3_{C_1} = 3

(b) In case of with replacement, A and B are independent.

So, the required probability is

$$P(A \cap B) = P(A) \times P(B)$$

$$= \frac{n(A)}{n(S)} \times \frac{n(B)}{n(S)}$$

$$= \frac{4_{C_1}}{10_{C_1}} \times \frac{4_{C_1}}{10_{c_1}}$$

$$= \frac{4}{10} \times \frac{4}{10}$$

$$= \frac{4}{25}$$

where,

n(A) = the number of outcomes that the 1st black marble is drawn from 4 black marbles

$= 4_{C_1} = 4$

n(B) = the number of outcomes that the 2nd black marble is also drawn from 4 black marbles

$= 4_{C_1} = 4$

(iii) Let, A = the 1st marble drawn is a white

B = the 2nd marble drawn is black

To find P(A∩B)

(a) In case of without replacement, A and B are not independent events.

So, the required probability is

P(A∩B)=P(A) × P(B|A)

$$= \frac{n(A)}{n(S_1)} \times \frac{n(B|A)}{n(S_2)}$$

$$= \frac{6_{c_1}}{10_{c_1}} \times \frac{4_{c_1}}{9_{c_1}}$$

$$= \frac{6}{10} \times \frac{4}{9}$$

$$= \frac{4}{15}$$

where,

n(A) = the number of outcomes that the 1st white marble is drawn from 6 white marbles

$= 6_{C_1} = 6$

n(B|A) = the number of outcomes that the 2nd black marble is drawn from the 4 black marbles

$= 4_{C_1} = 4$

(b) In case of with replacement, A and B are independent.

So, the required probability is

$$P(A \cap B) = P(A) \times P(B)$$

$$= \frac{n(A)}{n(S)} \times \frac{n(B)}{n(S)}$$

$$= \frac{6_{C_1}}{10_{C_1}} \times \frac{4_{C_1}}{10_{c_1}}$$

$$= \frac{6}{10} \times \frac{4}{10}$$

$$= \frac{6}{25}$$

Where,

n(A) = the number of outcomes that the 1st white marble is drawn from 6 white marbles

$= 6_{C_1} = 6$

n(B) = the number of outcomes that the 2nd black marble is also drawn from 4 black marbles

$= 4_{C_1} = 4$

Ex.9. A sum is given to the five students A, B, C, D and E. Their respective chances of solving it are$\frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \frac{1}{4}$ $\boldsymbol{and}\frac{1}{5}$. What is the probability that at least one of the students solves the sum?

Sol.

P [A can solve the sum] = $\frac{1}{2}$

⇨ P[A cannot solve the sum] = $1 - \frac{1}{2} = \frac{1}{2}$

Similarly,

P [B cannot solve the sum] = $1 - \frac{1}{3} = \frac{2}{3}$

P[C cannot solve the sum] = $1 - \frac{1}{4} = \frac{3}{4}$

P [D cannot solve the sum] = $1 - \frac{1}{4} = \frac{3}{4}$ and

P [E cannot solve the sum] = $1 - \frac{1}{5} = \frac{4}{5}$

So, P [no one can solve the sum] = $\frac{1}{2} \times \frac{2}{3} \times \frac{3}{4} \times \frac{3}{4} \times \frac{4}{5} = \frac{3}{20}$

Hence, P [at least one student solve the sum] = $1-\frac{3}{20}=\frac{17}{20}$

4.4. Conditional Probability and Baye's Theorem:

Conditional probability

Let us consider two events A and B.

The probability of occurring the event B given that the event A has already occurred is called the conditional probability of B given A. It is denoted by P(B|A) and is given by

$P(B|A)=\frac{P(A\cap B)}{P(A)}, P(A)>0$

The probability of occurring the event A given that the event B has already occurred is called the conditional probability of A given B. It is denoted by P(A|B) and is given by

$P(A|B)=\frac{P(A\cap B)}{P(B)}, P(B)>0$

Baye's theorem

Let, $A_1, A_2, \dots, A_k$ be the exhaustive, mutually exclusive and equally likely events in a sample space S i.e., $A_1 \cup A_2 \dots \cup A_k = S.$ Let, B be an event in S such that the happening of the event B depends on the happening of any one of the events $A_i; i=1,2, \dots k.$

Then the conditional event of A_i given B, is

$$\boldsymbol{P(A_i|B)=\frac{P(A_i)\times P(B|A_i)}{\sum P(A_i)\times P(B|A_i)}}$$

where, $P(A_i)$'s are equally likely probabilities of A_i.

$P(B|A_i)$'s are priori conditional probabilities of B given A_i

$P(A_i|B)$'s are posteriori conditional probabilities of A_i given B.

Ex.10. In a bolt factory machines M_1, M_2 and M_3 manufacture respectively 25, 35 and 45 percent of the total. Of their outputs 5, 4 and 2 percent respectively are defective bolts. One bolt is drawn at random from the product and is found defective. What is the probability that it was manufactured by the machine M_1?

Sol.

Let, A_1, A_2 and A_3 be the events that the bolts manufactured by the machines M_1, M_2 and M_3 respectively.

A/Q, $P(A_1) = 0.25, P(A_2) = 0.35$ and $P(A_3) = 0.45$

Let, B be the event that a bolt selected at random, is defective.
A/Q, $P(B|A_1) = 0.05, P(B|A_2) = 0.04$ and
$P(B|A_3) = 0.02$

The required probability is
P(Randomly selected bolt was manufactured by the machine M1 given that it is defective)

$$= P(A_1|B) = \frac{P(A_1)\times P(B|A_1)}{\sum P(A_i)\times P(B|A_i)}$$

$$= \frac{0.25\times 0.05}{0.25\times 0.05+0.35\times 0.04+0.45\times 0.02} = \frac{0.0125}{0.0355} = 0.3521$$

Ex.11. A black ball is drawn from one of the three bags. The first contains three black balls and seven white balls, the second contains five black and three white balls and the third contains eight black and four white balls. What is the probability that it was drawn from the second bag?

Sol.

Given,

First bag	: 3 black + 7 white = 10 balls
Second bag	: 5 black + 3 white = 8 balls
Third bag	: 8 black + 4 white = 12 balls

Let, A_1, A_2 and A_3 be the events of the selection of the first, second and third bags respectively.

Since, the selection of the bags are equally likely, so, $P(A_1) = P(A_2) = P(A_3) = \frac{1}{3}$

Let, the event B denotes that a black ball is drawn from a bag.

$P(B|A_1) = \frac{3}{10}, P(B|A_2) = \frac{5}{8}$, and $P(B|A_3) = \frac{8}{12} = \frac{2}{3}$

The required probability is P(a ball was drawn at random from the second bag given that it is black) $= P(A_2|B)$

$= \frac{P(A_2)\times P(B|A_2)}{\sum P(A_i)\times P(B|A_i)}$

$= \frac{\frac{1}{3}\times\frac{5}{8}}{\frac{1}{3}\times\frac{3}{10}+\frac{1}{3}\times\frac{5}{8}+\frac{1}{3}\times\frac{2}{3}}$

$= \frac{\frac{5}{24}}{\frac{191}{360}}$

$= \frac{75}{191} = 0.3927$

Ex.12. The probability that a football player will play on ordinary ground is 0.6 and on green turf is 0.4. The probability that he will get knee injury when play on ordinary ground is 0.07 and that a green turf is 0.04. What is the probability that he got knee injury due to play on ordinary ground?

Sol.

Let, the events A_1 and A_2 be that the player will play on ordinary ground and on green turf respectively.

A/Q, $P(A_1) = 0.6$ and $P(A_2) = 0.4$

Let, B be the event that the player will get knee injury when
play on a ground.

A/Q, $P(B|A_1) = 0.07$ and $P(B|A_2) = 0.04$

The required probability is P(he got knee injury due to play on ordinary ground)

$= P(A_1|B)$

$= \frac{P(A_1)\times P(B|A_1)}{\sum P(A_i)\times PB|A_i)}$

$= \frac{0.6\times0.07}{0.6\times0.07+0.4\times0.04}$

$= \frac{0.042}{0.058} = 0.7241$

4.5. Random variable and Probability distribution

Random Variable (RV): A random variable (RV) is a function defined on a sample space of real values. In other words, a random variable is a quantity which takes a value in a sample space.

Types of random variable: There are two types of random variable viz.

(i) Discrete or discontinuous random variable and

(ii) Continuous random variable.

The random variable which takes a specified countable or uncountable real values of a sample space, is called **discrete random variable.**

For example: If the random variable X be the number of heads obtained in tossing of two coins, then X = {0, 1, 2}. In a family of four children, if the random variable X be the number of girls, then X = {0, 1, 2, 3, 4}.

In both the cases, the values of the random variable are countable and specified.

If the random variable X be the number of bacteria formed in a one gram of curd, then X = {1, 2, 3, …}. The values of X are uncountable and specified.

The random variable which takes any real value of a sample space of uncountable real numbers or any real value within two real numbers, is called **continuous random variable**. Thus, if X be a continuous random variable, then $-\infty < X < \infty$ or $a < X < b\ or\ a \leq X \leq b.$

For example: Day temperature, heights of students in a class, marks obtained by students in a certain subject etc.

Probability distribution: The distribution of all the values of a random variable X with their corresponding probabilities, P(x) or f(x), is known as probability distribution of the random variable X.

Probability distribution is of two types – **discrete probability distribution and continuous probability distribution**.

In discrete probability distribution, the underlying variable X is a discrete random variable and in continuous probability distribution, the underlying variable X is a continuous random variable. Binomial and Poisson distributions are discrete probability distribution, while. Normal distribution is a continuous probability distribution.

4.6. Mathematical Expectation and Variance of a random variable

Mathematical Expectation of a random variable

Let, X be a discrete random variable taking the values $x_1, x_2, \dots x_n$ with corresponding probabilities P(x1), P(x2),…,P(xn). Then the mathematical expectation of X is given by

$$E(X) = x_1 \times P(x_1) + x_2 \times P(x_2) + \cdots + x_n \times P(x_n)$$
$$= \sum xP(x)$$

Let, X be a continuous random variable taking the values in $(-\infty, \infty)$, i.e., $-\infty < X < \infty$ with probability f(x) (say), then the mathematical expectation of X is given by

$E(X) = \int_{-\infty}^{\infty} xf(x)dx$ [out of the module]

Properties of Mathematical Expectation

(i) E (C) = C, where, C is a constant.

(ii) E(X) = $\bar{X}, the\ mean\ of\ X$

(iii) $E(X - \bar{X}) = 0$

(iv) E (CX) = $C \times E(X)$, where, C is a constant

(v) $E(aX \pm b) = a \times E(X) \pm b$, where, a and b are constants.

(vi) If $X_1, X_2, \dots, X_n$ be the random variables and $a_1, a_2, \dots a_n$, be their corresponding coefficients, then $E(a_1X_1 \pm a_2X_2 \pm \cdots \pm a_nX_n) = a_1E(X_1) \pm a_2E(X_2) \pm \cdots \pm a_nE(X_n)$

Variance of a random variable

The variance of a random variable X, is defined as $\boldsymbol{Var(X) or\ \sigma_X^2 = E[X - E(X)]^2}$

Or $\boldsymbol{Var(X) = E(X^2) - [E(X)]^2}$

Properties of variance

(i) $Var(C) = 0$, where, C is a constant

(ii) $Var(X) \geq 0$

(iii) $Var(CX) = C^2Var(X)$, where, C is a constant.

(iv) $Var(aX \pm b) = a^2 var(X)$, where, a and b are constants

(v) If $X_1, X_2, \ldots, X_n$ be the random variables and $a_1, a_2, \ldots a_n$, be their corresponding coefficients, then $Var(a_1X_1 \pm a_2X_2 \pm \cdots \pm a_nX_n)$

$$= a_1^2 Var(X_1) \pm a_2^2\, Var(X_2) \pm \cdots \pm a_n^2 Var(X_n)$$

Ex.13. Three unbiased coins are tossed at random, find the expectation and variance of getting head.

Sol.

If three unbiased coins are tossed, then the sample space is given by

$S = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$

Let, X = event of getting number of heads = {0, 1, 2, 3}

∴ P(X = 0)

= P(no head obtained)

= P(out of 8 outcomes one may occur which is TTT)

$= \frac{1}{8}$

P(X = 1)

= P(one head obtained)

= P(out of 8 outcomes 3 outcomes may occur, which are HTT, THT, TTH)

$= \frac{3}{8}$

P(X = 2)

= P(two heads obtained)

= P(out of 8 outcomes 3 outcomes may occur, which are HHT, HTH, THH)

$= \frac{3}{8}$

P(X = 3)

= P(three heads obtained)

= P(out 8 outcomes one may occur, which is HHH)

$= \frac{1}{8}$

$\boldsymbol{X = x}$	$\boldsymbol{P(x)}$	$\boldsymbol{x \times P(x)}$	$\boldsymbol{x^2 \times P(x)}$
$X = 0$	$\frac{1}{8}$	$0 \times \frac{1}{8} = 0$	$0 \times 0 = 0$

$X = 1$	$\frac{3}{8}$	$1 \times \frac{3}{8} = \frac{3}{8}$	$1 \times \frac{3}{8} = \frac{3}{8}$
$X = 2$	$\frac{3}{8}$	$2 \times \frac{3}{8} = \frac{6}{8}$	$2 \times \frac{6}{8} = \frac{12}{8}$
$X = 3$	$\frac{1}{8}$	$3 \times \frac{1}{8} = \frac{3}{8}$	$3 \times \frac{3}{8} = \frac{9}{8}$
Total	$\boldsymbol{\sum P(x)}$ $\mathbf{= 1}$	$\boldsymbol{\sum xP(x)}$ $\mathbf{= \frac{12}{8}}$ $\mathbf{= 1.5}$	$\boldsymbol{\sum x^2 P(x)}$ $\mathbf{= \frac{24}{8} = 3}$

$E(X) = \sum xP(x)$

$= 1.5$

That is, the expectation of getting head is 1.5

$$Var(X) = E(X^2) - [E(X)]^2$$

$$=> Var(X) = \sum x^2 P(x) - 1.5^2$$

$$= 3 - 2.25 = 0.75$$

That is, the variance of getting head is 0.75

Check your progression

Ex.14. Two unbiased coins are tossed at random, find the expectation and variance of getting head. [Answer: E(X) = 1 and Var (X) = 0.75]

Ex.15. A fair dice is thrown at random, what is the expectation and variance of number obtained.

Sol.

A dice is thrown at random, S = {1, 2, 3, 4, 5, 6}

Let, X be a number obtained when a fair dice is thrown.

So, X = {1, 2, 3, 4, 5, 6}

$X = x$	$P(x)$	$xP(x)$	$x^2P(x)$
$X = 1$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

$X=2$	$\frac{1}{6}$	$\frac{2}{6}$	$\frac{4}{6}$
$X=3$	$\frac{1}{6}$	$\frac{3}{6}$	$\frac{9}{6}$
$X=4$	$\frac{1}{6}$	$\frac{4}{6}$	$\frac{16}{6}$
$X=5$	$\frac{1}{6}$	$\frac{5}{6}$	$\frac{25}{6}$
$X=6$	$\frac{1}{6}$	$\frac{6}{6}$	$\frac{36}{6}$
Total	$\sum P(x)$ $=1$	$\sum xP(x)$ $=\frac{21}{6}$ $=3.5$	$\sum x^2P(x)$ $=\frac{91}{6}$ $=15.17$

E(X) = $\sum xP(x) = 3.5$

and Var (X) =$E(X^2) - [E(X)]^2$

$= \sum x^2P(x) - 3.5^2$

$= 15.17 - 12.25 = 2.92$

4.7 Probability distribution: Binomial distribution, Poisson distribution and Normal distribution

Binomial distribution

Binomial distribution is a discrete probability distribution based on Bernoulli's trial, i.e, the experiment having two possible outcomes.

Definition:

A discrete random variable X (say) is said to follow a binomial distribution with parameters 'n' (the number of independent trials) and the probability of a success 'p' (say) in single trial, if its probability distribution is given by

$$\boldsymbol{P(x) = n_{C_x}p^x(1-p)^{n-x}; x = 0, 1, 2, 3, \dots, n}$$

Assumptions of Binomial Distribution

(i) The trials are independent and the number of trials (n) is finite.

(ii) The probability of a success (p) in each toss is equal.

(iii)The sum of the probability of a success (p) and the probability of a failure (1 – p) is 1.

(iv) The possible outcomes in each trial is same.

Properties of Binomial Distribution

(i) Binomial distribution is a discrete probability distribution with parameters ‘n’ (say) (the number of independent trials) and ‘p’ (say) (the probability of a success).

(ii) The mean of the binomial variate X is

$\boldsymbol{\mu_X}$ =**E(X) = np** and

the variance is $\boldsymbol{\sigma_X^2}$ =**Var (X) = np (1 – p)**

So, **mean** > **variance** and the standard deviation is $\boldsymbol{\sigma_X = \sqrt{np(1-p)}}$

(iii)The binomial variate X takes the values say 0, 1, 2, …, n and the total probability for all the values of X is 1. That is, $\boldsymbol{P(0) + P(1) + P(2) + \cdots + P(n) = 1}$

(iv)The mode of this distribution is that value of X, which has the maximum probability.

Examples of Binomial Distribution

(i) To find the number of defective items in a sample in a manufacturing process, if the probability of a defective item is not small.

(ii) To find the number of students passed in a class test.

(iii)The number of male babies born on a particular period at a health center.

Ex.16. If the probability of male birth is 0.5, find the probability that in a family of four children, there will be

(i) At least one boy

(ii) Two boys and two girls

(iii)**One girl.**

Sol.

Let X be the number of boys in a family of 4 children

The probability of a male birth in a family is p = 0.5

Therefore, $X \sim Binomial(n = 4, p = 0.5)$

So, $P(x) = n_{C_x} p^x (1-p)^{n-x}; x = 0,1,2,3, \dots, n$

$=> P(x) = 4_{C_x} \times 0.5^x \times 0.5^{4-x}, x = 0,1,2,3,4$

(i) P (at least one boy)

= $P(X \geq 1)$

$= P(1) + P(2) + P(3) + P(4)$

$= 1 - P(0) \qquad [\because P(0) + P(1) + P(2) + P(3) + P(4) = 1]$

$= 1 - 4_{C_0} \times 0.5^0 \times 0.5^4$

$= 1 - 1 \times 1 \times 0.0625$

$= 1 - 0.0625$

$= 0.9375$

(ii) P (two boys and two girls)

= P (two boys)

$= P(2)$

$= 4_{C_2} \times 0.5^2 \times 0.5^2$

$= \frac{4\times3}{2\times1} \times 0.25 \times 0.25$

$= 0.375$

(iii) P (one girl)

= P(3 boys)

=P(3)

=$4_{C_3} \times 0.5^3 \times 0.5$

$= \frac{4\times3\times2}{3\times2\times1} \times 0.125 \times 0.5$

$= 0.25$

Ex.17. A machine produces on average 10% defective items. Find the probability that in a sample of 10 items chosen at random, (i) at least one will be defective (ii) exactly three items are defective, (iii) less than three items are defectives.

Sol.

Let X be the number of defective items produced in a sample of 10 items.

The probability of a defective item is p = 0.1

Therefore, $X \sim Binomial(n = 10, p = 0.1)$

So, $P(x) = n_{C_x} p^x (1-p)^{n-x}; x = 0,1,2,3, \dots, n$

⇨ $P(x) = 10_{C_x} \times 0.1^x \times 0.9^{10-x}, x = 0,1,2, \dots, 10$

(i) $P(at\ least\ 1\ will\ be\ defective)$

$= P(X \geq 1)$

$= P(1) + P(2) + \cdots + P(10)$

$= 1 - P(0)\ \ [\because P(0) + P(1) + P(2) + \cdots + P(10) = 1]$

$= 1 - 10_{C_0} \times 0.1^0 \times 0.9^{10}$

$= 1 - 0.3487$

$= 0.6513$

(ii)$P(exactly\ 3\ items\ are\ defectives)$

$= P(3)$

$= \frac{10\times9\times8}{3\times2\times1} \times 0.001 \times 0.4783$

$= 0.0574$

(iii) $P(less\ than\ 3\ items\ are\ defectives)$

$= P(X < 3)$

$= P(0) + P(1) + P(2)$

$= 0.3487 + 10_{C_1} \times 0.1^1 \times 0.9^9 + 10_{C_2} \times 0.1^2 \times 0.9^8$

= 0.3487 + 10 × 0.1 × 0.9^9+ $\frac{10\times9}{2\times1}$ × 0.01 × 0.9^8

= 0.3847 + 0.3874 + 0.1937

= 0.9658

Check your progress

Ex.18. If 20% of the items produced by a machine are defectives, find the probability that out of 4 bolts chosen at random at most 2 items will be defectives. [Answer: $P(X \leq 2) = 0.9728$]

Ex.19. If the probability of recovering loan amount according to repayment schedule is 0.8 for a particular category of loans, what is the probability of recovering at least 4 out of 6 loans sanctioned in this category? Also calculate the expected number of recoveries and extent of variation.

Sol.

Let, X be the number of recoveries out of 6 loans sanctioned.

Probability of a recovery loan amount is p = 0.8

So, $X \sim Binomial(n = 6, p = 0.8)$

Therefore,

$P(x) = n_{C_x} \times p^x \times (1-p)^{n-x}; x = 0,1,2,3, \dots, n$

$=> P(x) = 6_{C_x} \times 0.8^x \times 0.2^{6-x}; x = 0,1,2,3,4,5,6$

The required probability is $P(recovering\ at\ least\ 4\ out\ of\ 6)$

$= P(X \geq 4)$

= P (4) + P (5) + P (6)

= $6_{C_4} \times 0.8^4 \times 0.2^2 + 6_{C_5} \times 0.8^5 \times 0.2^1 + 6_{C_6} \times 0.8^6 \times 0.2^0$

$= 6_{C_2} \times 0.4096 \times 0.04 + 6_{C_1} \times 0.32768 \times 0.2 + 6_{C_0} \times 0.262144 \times 1$

$[since, n_{c_r} = n_{c_{n-r}}]$

$= \frac{6\times5}{2\times1} \times 0.4096 \times 0.04 + 6 \times 0.32768 \times 0.2 + 1 \times 0.262144$

$= 0.24576 + 0.393216 + 0.262144$

$= 0.9011$

Expected number of recoveries is μ_X= np = 6 × 0.8 = 4.8

Extent of variation $\sigma_X = \sqrt{np(1-p)}$

$$= \sqrt{6 \times 0.8 \times 0.2} = 0.98$$

Ex.20. Bring out the fallacy in the following statement

“The mean of a binomial distribution is 10 and standard deviation is 6”

Ans:

The mean of a binomial distribution is 10

The standard deviation is 6

So, the variance is $6^2 = 36$

∴ Mean < Variance

So, the given statement is wrong, since, in a binomial distribution mean > variance

Ex.21. A student obtained the following results. Is this result consistent?
For a binomial distribution, mean = 4, variance = 6.

Ans:

By the question, mean < variance

∴ The result is not consistent, since, in a binomial distribution mean > variance.

Poisson Distribution

A binomial distribution will tend to become a Poisson distribution, if

(i) The number of independent trials is very large, i.e, $n \to \infty$

(ii) The probability of a success is very small, i.e, $p \to 0$

(iii) $n \times p = \lambda$(constant)

Definition of Poisson Distribution

A discrete random variable X is said to follow a Poisson distribution with parameter λ, if the probability distribution of X, is given by $\boldsymbol{P(x) = \frac{e^{-\lambda}\times\lambda^{x}}{x!}, x = 0, 1, 2, 3, \ldots and\ e = 2.7183(approx)}$

Properties of Poisson Distribution

(i) Poisson distribution is a discrete probability distribution with parameter λ

(ii) Mean of a Poisson variate X, is μ_X =E (X) = λ and the variance is σ_X^2 =Var (X) = λ. So, **mean = variance** and the standard deviation is $\sqrt{\lambda}$

(iii)The random variable of this distribution say X takes the values 0, 1, 2, … and the total probability for all the values of X, is 1. That is, $P(0) + P(1) + P(2) + \cdots = 1$

(iv)The mode of this distribution is that value of the variable X, whose corresponding probability is maximum.

Examples of Poisson Distribution

Poisson distribution is used those events, whose the chances of occurring are very less. So, some of the physical situations of this distribution are given below:

(i) To find the number of defective items occurred in well known manufacturing process.

(ii) To find the number the number mistakes or error found in a text book published by a well known publishers.

(iii)To find the number of errors found in the pass book of a customer.

(iv)To find the number of persons died due to rare disease in a City.

Ex.21. If 1% of the bolts produced by a certain machine are defectives, find the probability that in a random sample of 300 bolts,

(i) All bolts are good

(ii) Exactly 2 bolts are defectives

(iii) At least 1 bolt is defective.

Sol.

Here, n = 300, which is considered as large

The probability of a defective bolt is p = 0.01, which is small.

Let, X be the number of defective bolts in a sample of 300 bolts

So, $X \sim Poisson(\lambda)$, where, λ= $n \times p = 300 \times 0.01 = 3$

Therefore, $P(x) = \frac{e^{-\lambda} \times \lambda^x}{x!}, x = 0,1,2,3, \dots and\ e = 2.7183$

⇨ $P(x) = \frac{e^{-3} \times 3^x}{x!}, x = 0,1,2,3, \dots$

$= \frac{0.0498 \times 3^x}{x!}, x = 0,1,2,3,\dots$

(i) P [all bolts are good]

= P [all bolts are non-defectives]

= P [X = 0]

$= P(0)$

$= \frac{0.0498 \times 3^0}{0!}$

$= \frac{0.0498 \times 1}{1} = 0.0498$

(ii) P [exactly 2 bolts are defectives]

= P [X = 2]

= $P\ (2)$

$= \frac{0.0498 \times 3^2}{2!}$

$= \frac{0.4482}{2} = 0.2241$

(iii) P [at least 1 bolt is defective]

$= P[X \geq 1]$

$= P[X = 1] + P[X = 2] + \cdots$

$= 1 - P[X = 0]$

$= 1 - \frac{0.0498 \times 3^0}{0!}$

= 1 – 0.0498 = 0.9502

Check your progress

Ex.22. If 2% electric bulbs manufactured by a certain company are defectives, find the probability that in a sample of 200 bulbs

(i) Less than 2 bulbs are defectives

(ii) More than 3 bulbs are defectives [Given, $e^{-4} = 0.0183$]

[Answer: (i) P(X<0) = 0.0916, P(X>3) = 0.5665]

Ex.23. A manufacturer of copper pins knows that 2% of his product is defective. If he sells copper pins in boxes of 200 guarantees that not more than 5 pins will be defective, what is the probability that a box will fail to meet the guaranteed quality? [Given, $e^{-4} = 0.0183$]

Sol.

Here, n = 200, which is large

The probability of a defective pin p = 0.02, which is small

So, $X \sim Poisson(\lambda)$, where, $\lambda = n \times p = 200 \times 0.02 = 4$

Therefore,

$P(x) = \frac{e^{-\lambda} \times \lambda^x}{x!}, x = 0,1,2,3, \ldots and\ e = 2.7183$

⇨ $P(x) = \frac{e^{-4} \times 4^x}{x!}$

$= \frac{0.0183 \times 4^x}{x!}; x = 0,1,2,3, \ldots$

P [the box fail to meet guarantee]

= P [more than 5 pins will be defectives]

$=P[X > 5]$

$= P[X = 6] + P[X = 7] + \cdots$

$= 1 - P[X = 0] - P[X = 1] - P[X = 2] - P[X = 3] - P[X = 4] - P[X = 5]$

$= 1 - 0.0183 - 0.0732 - 0.1464 - 0.1952 - 0.1952 - 0.15616$

$= 0.2155$

Normal / Gaussian Distribution:

A binomial variate X (say) with parameters ‘n’ and ‘p’ will tend to become a normal distribution if

(i) The number of independent trials is very large i.e, $n \to \infty$

(ii) The probability of a success is not small.

Definition:

A continuous random variable X is said to follow a normal distribution with parameters μ and σ, if its probability distribution is given by

$$\boldsymbol{f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, -\infty < x < \infty}$$

Where, μ = E (X), the mean or expectation of X

σ = standard deviation of X = $\sqrt{Var(X)}$

π = $\frac{22}{7} = 3.14286(approx)$

e = 2.71828

Standard Normal Variate (SNV)

If $X \sim N(\mu, \sigma)$, then we define a random variable Z, such that $\boldsymbol{Z = \frac{X-\mu}{\sigma}}$, which follows a normal distribution with mean 0 and standard deviation 1. In short, $\boldsymbol{Z \sim N(0, 1)}$. Z is called standard normal variate and the pdf of Z is

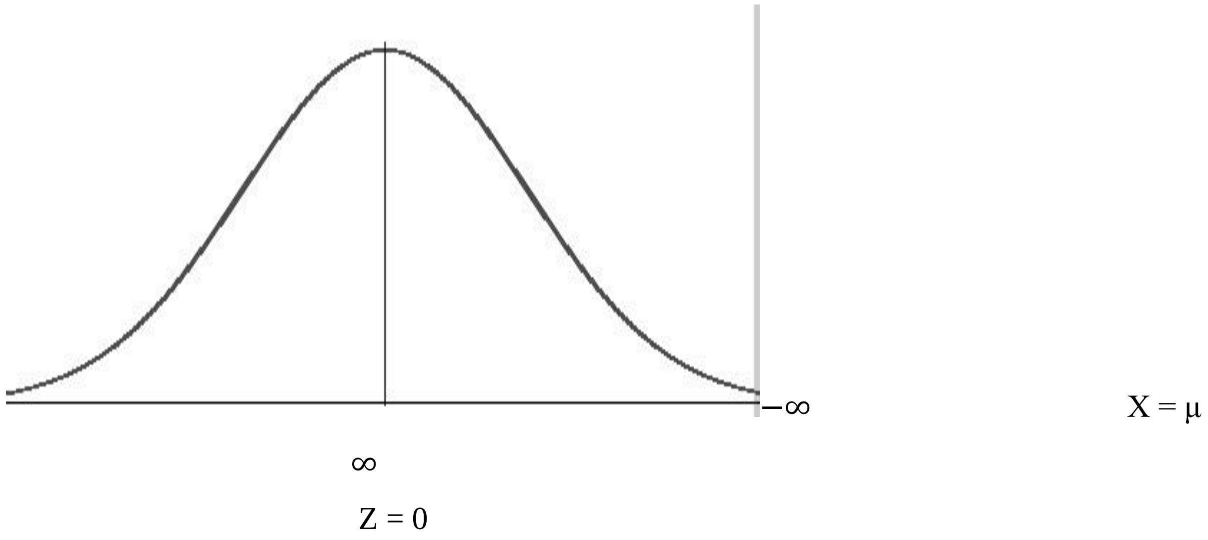
$$\boldsymbol{f_Z(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}, -\infty < z < \infty}$$

Properties of Normal Distribution

(i) It is a continuous probability distribution with mean μ and standard deviation σ. Its probability density function is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, -\infty < x < \infty$$

(ii) The curve is bell shaped and symmetrical about the line X = μ



(iii)Mean, median and mode lies at the same point i.e, mean = median = mode

(iv)$f(x)$ decreases rapidly as x increases

(v) The maximum probability occurs at the point X = μ and is $\frac{1}{\sigma\sqrt{2\pi}}$

(vi)Mean deviation about mean = $\frac{4}{5}\sigma$

(vii) Since f(x), being the probability, can never be the negative, so that no portion of the curve lies below the X-axis.

(viii) Area properties

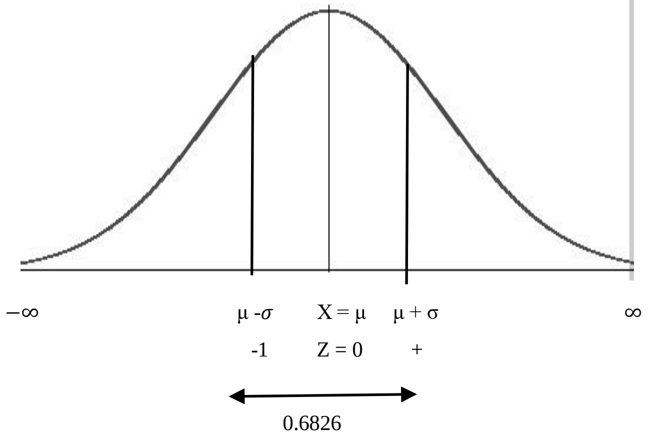
Total area of the normal curve is 1 i.e, $P(-\infty < X < \infty) = 1$

$\therefore P(-\infty < X < \mu) = P(\mu < X < \infty) = 0.50$

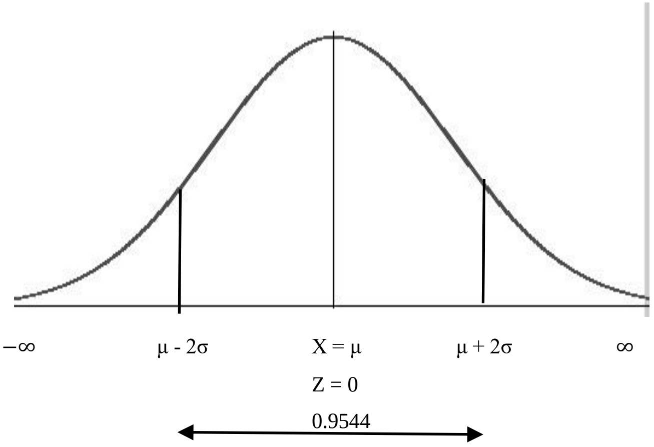
$or\ P(-\infty < Z < 0) = P(0 < Z < \infty) = 0.50$

(i) The area of the normal curve between μ – σ and μ + σ is 0.6826

That is, P (μ – σ < X < μ + σ) = P (-1 < Z < 1) = 0.6826

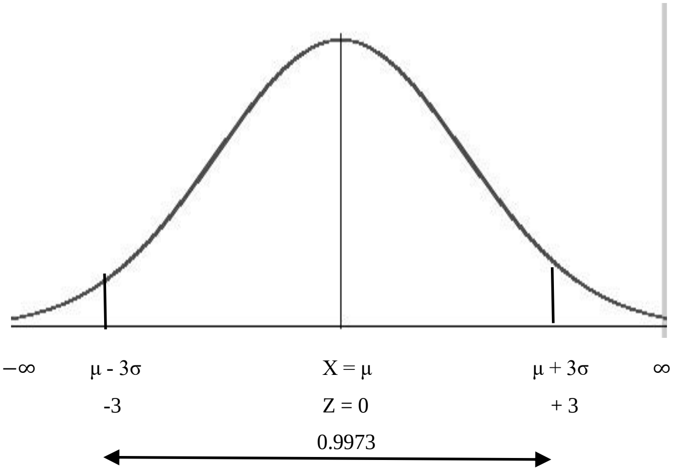


(ii) The area of the normal curve between μ – 2σ and μ + 2σ is 0.9544. That is, P (μ – 2σ < X < μ + 2σ) = P (- 2 < Z < 2) = 0.9544



(iii) The area of the normal curve between μ – 3σ and μ + 3σ is 0.9973

That is, P (μ – 3σ < X < μ + 3σ) = P (- 3 < Z < 3) = 0.9973



Ex.24. The distribution of monthly income of 500 workers be assumed to be normal with mean ₹2000 and standard deviation of ₹200. Estimate the number of workers with incomes:-

- **(i) Exceeding ₹2300 pm**
- **(ii) Between ₹1800 pm and ₹2300 pm**
- **(iii) Below ₹1800 pm**

[Given Z = 0.67 1.00 1.5

A = 0.75 0.8413 0.9322]

Sol.

Let, X be the incomes of workers per month.

A/Q, $X \sim N(\mu = 2000, \sigma = 200)$

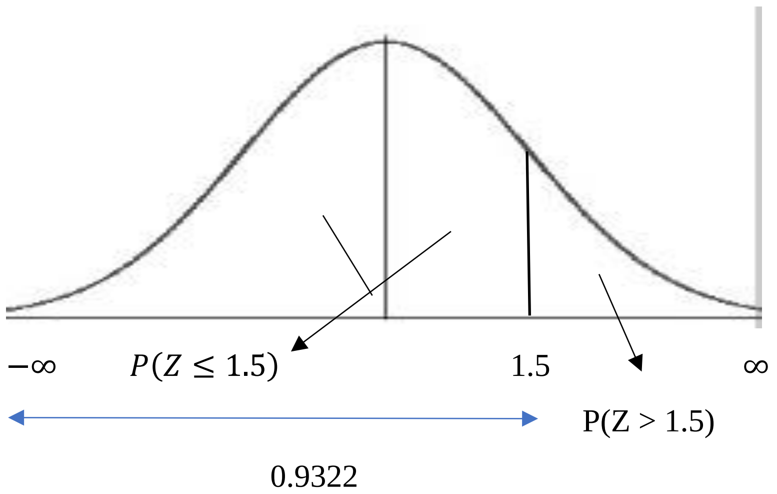
(i) $P(X > 2300) = P\left(Z > \frac{2300-\mu}{\sigma}\right)$

$$= P\left(Z > \frac{2300-2000}{200}\right)$$

$$= P(Z > 1.5)$$

$$= 1 - P(Z \leq 1.5)$$

$= 1 - 0.9322 = 0.0678$



Therefore, the estimated number of workers with incomes exceeding ₹2300 pm out of 500 workers is

500 × 0.0678

= 33.9 ≅ 34

(ii) $P(1800 < X < 2300)$

$$= P\left(\frac{1800-\mu}{\sigma} < Z < \frac{2300-\mu}{\sigma}\right)$$

$$= P\left(\frac{1800-2000}{200} < Z < \frac{2300-2000}{200}\right)$$

$$= P(-1 < Z < 1.5)$$

$= P(-1 < Z < 0) + P(0 < Z < 1.5)$

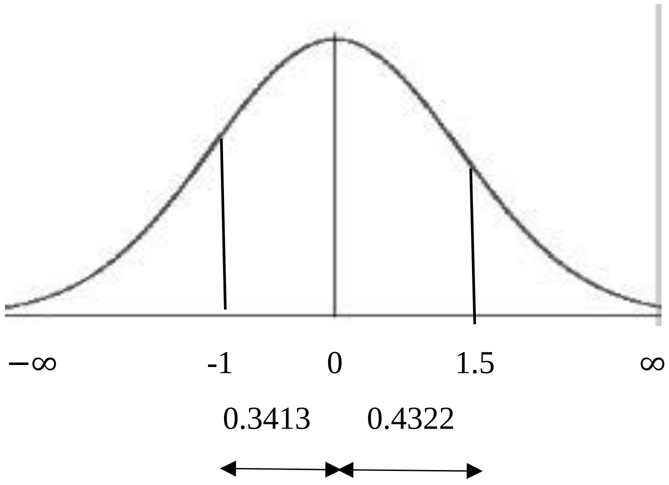
= (Area between -1 and ∞ - area between 0 and ∞)

+ (Area between -∞ and 1.5 - area between -∞ and 0)

= (0.8413 – 0.50) + (0.9322 – 0.50)

= 0.3413 + 0.4322 = 0.7735

Therefore, the estimated of workers with incomes between ₹1800 and ₹2300 out of 500 workers is $500 \times 0.7735 = 386.75 \cong 387$



(iii)$P(X < 1800)$

$= P\left(Z < \frac{1800-\mu}{\sigma}\right)$

$= P\left(Z < \frac{1800-2000}{200}\right)$

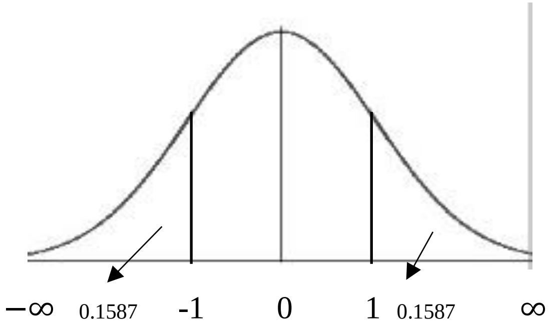
$= P(Z < -1)$

$= P(Z > 1)$

$= 1 - P(Z \leq 1)$

$= 1 - 0.8413 = 0.1587$

Therefore, the estimated number of workers with incomes below ₹1800 pm out of 500 workers is $500 \times 0.1587 = 79.35 \cong 79$



Check your progress

Ex.25. If the heights of 500 students are normally distributed with mean 68.0 inches and standard deviation 3.0 inches, how many students have height

(i) Greater than 72 inches

(ii) Less than 64 inches

(iii) Between 65 and 71 inches

[Given, Z= 1.00 1.33
A= 0.8413 0.9082]

[Answer: (i) $P(X>72)=0.0918$, so, 46 students

(ii) $P(X<64)=0.0918$, so, 46 students

(iii)$P(65<X<71)=0.6826$ so, 341 students]

Importance of Normal Distribution

(i) Data obtained from psychological, physical and biological measurements approximately follow Norma distribution.

(ii) Distribution like Binomial, Poisson etc. can be approximated to normal distribution.

(iii)Normal curve is used to find the confidence limits of the population parameters.

(iv)Normal distributions are largely applied in Statistical Quality Control (SQC) in Industry for finding control limits.

(v) The theory of errors of observations in physical measurements are based on normal distribution.

(vi)For large samples, any statistic (sample mean, sample variance etc.) follows normal distribution and as such it can be studied with the help of normal curve.

-----xxxxx-----

UNIT 5: SAMPLING DISTRIBUTION, ESTIMATION (CONCEPT ONLY) AND TESTING OF HYPOTHESIS

Structure

5.1. Population, sample and sampling

Population or Universe: An aggregate of all facts or data under study, is called a population. A population is either finite or infinite.

For example: If we study the average life of an electric bulb manufactured by a Company, then all the bulbs manufactured by the Company, will constitute a population.

Sample: A part or fraction of a population is called a sample. A sample is always finite.

For example: If we study the average life of an electric bulb manufactured by a Company, then testing the average life of each bulb in the population is not feasible. In such case we select a few bulbs, which is called a sample for the testing.

Sampling: A method of drawing sample(s) from a population, is called a sampling. There are three types of sampling viz., **Probability or Random sampling and Non-probability or non-random sampling.**

5.2. Types of sampling

There are two types of sampling, viz., **probability or random sampling** and **non-probability or non-random sampling.**

In probability (random) sampling, all the units have a chance of being chosen for the sample. It can be estimated the effect of sampling error and hence there is a chance to be representative sample It is more time consuming and expensive than non-probability sampling.

In non-probability (non-random) sampling, the units have no chance of being selected. Consequently, it cannot be estimated the effect of sampling error and there is a significant risk to be non-representative sample which produces non-generalisable results. Non-probability sampling is cheaper and more convenient and they are useful for exploratory research.

5.3. Parameter and Statistic

Population characteristics are called **parameters**. In other words, population mean, population variance, population proportion etc. are known as parameters. The value of a parameter is constant for a particular population.

Sample characteristics are called **statistics.** In other words, sample mean, sample variance, sample proportion etc., are known as statistics. The value of a statistic is not a constant, it is a variable, since, it changes sample to sample.

Descriptive Statistics Vs Inferential Statistics

Descriptive Statistics: It describes the important characteristics/ properties of the data using the measures the central tendency like mean/ median/mode and the measures of dispersion like range, standard deviation, variance etc. Data can be summarized and represented in an accurate way using charts, tables and graphs.

For example: We have marks of 1000 students and we may be interested in the overall performance of those students and the distribution as well as the spread of marks. Descriptive statistics provides us the tools to define our data in a most understandable and appropriate way.

Inferential Statistics: It is about using data from sample and then making inferences about the larger population from which the sample is drawn. The goal of the inferential statistics is to draw conclusions from a sample and generalize them to the population. It determines the probability of the characteristics of the sample using probability theory. The most common methodologies used are hypothesis tests, Analysis of variance etc.

For example: Suppose we are interested in the exam marks of all the students in India. But it is not feasible to measure the exam marks of all the students in India. So now we will measure the marks of a smaller sample of students, for example 1000 students. This sample will now represent the large population of Indian students. We would consider this sample for our statistical study for studying the population from which it's deduced.

Difference between Descriptive Statistics and Inferential Statistics

Descriptive Statistics	**Inferential Statistics**
To describe the target population	To make inference from the sample and generalise it to the population
Organise, analyse and present the data in a meaningful manner.	Compare, test and predicts the future outcomes.
The results are shown in charts, tables and graphs	The results are shown in terms of probability.
Describes only known data	Tries to make conclusion

	about the data which are beyond the sample, but included in the population.
Tools used are Measures of central tendency, dispersion, skewness, kurtosis, correlation and regression etc.	Tools used are Estimation and Testing of Hypothesis.

5.4. Sampling distribution of a statistic

Let, k samples are drawn from a population. If t be a statistic, let t_1, t_2, t_3,…,t_k be the values of the statistic t. Then a distribution of all the values of t with their corresponding frequencies, is called the sampling distribution of the statistic t. Following are the measures of the sampling distribution of the statistic t.

Since, the statistic 't' is a variable, so, mean and variance of the sampling distribution of the statistic 't' are determined by using the following formulae

Mean of t,

$$\bar{t} = \frac{t_1+t_2+\cdots+t_k}{k} = \frac{\sum t}{k} \quad \text{[in case of ungrouped data]}$$

$$= \frac{f_1t_1+f_2t_2+\cdots+f_kt_k}{f_1+f_2+\cdots+f_k} = \frac{\sum ft}{\sum f} \quad \text{[in case of grouped data]}$$

Variance of t,

$$\sigma_t^2 = \frac{\sum(t-\bar{t})^2}{k-1} \qquad \text{[in case of ungrouped data]}$$

$$= \frac{\sum f(t-\bar{t})^2}{\sum f-1} \qquad \text{[in case of grouped data]}$$

5.5. Standard error of a statistic

The standard deviation of the sampling distribution of the statistic t, is called the standard error of the statistic t. Thus, if σ_t^2 be the variance of the statistic t, then σ_t is the standard error of the statistic t.

For example, the standard error of the sample mean $\bar{X}$ is $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}, where, \sigma$ is the population standard deviation and n is the sample size.

Importance of Standard Error

(i) Standard error is used to estimate the confidence interval of an unknown parameter.

(ii) Standard error is used to test the statistical hypothesis.

(iii)Standard error is used to measure the precision of a statistic.

5.6. Methods of estimation (Concept only)

Estimation is a process of finding the value of an unknown parameter by using the sample values. Estimation has two types – **Point estimation** and **Interval estimation.**

Point Estimation

The process of finding the single value of an unknown parameter by using the sample values, is known as Point estimation. A point estimator is said to be a good estimator, if its value is very near to the true value of the unknown parameter. A good estimator has the following criteria.

- Unbiasedness
- Consistency
- Efficiency
- Sufficiency

Interval Estimation

Let, θ be a parameter, which is unknown and $\hat{\theta}$, the corresponding statistic. Let, $\hat{\theta}_1$ and $\hat{\theta}_2$ be the values of the statistic $\hat{\theta}$ so that $\hat{\theta}_1 < \hat{\theta}_2$. If $P\left[\hat{\theta}_1 < \theta < \hat{\theta}_2\right] = 1 - \alpha,$ then the limits $\hat{\theta}_1$ and $\hat{\theta}_2$ are called **interval estimation** or **confidence interval** of the parameter $\theta,$ where, $1 - \alpha$ is called confidence level or coefficient of confidence, the values of which are generally 0.90 or 0.95 or 0.99 and α is called level of significance, the values of which are generally 0.10 or 0.05 or 0.01.

Difference between Point estimation and Interval estimation

Point estimation	Interval estimation
1. It is a process of estimating single value of a parameter	1. It is a process of estimating two values of a statistic where, the true value of an unknown parameter lies.
2. A good estimator is that value of a statistic, which value is very near to the true value of the parameter.	2. As the sample size increases, the width of the interval estimation decreases at a certain confidence level and hence we get a good interval estimation.
3. Criteria of a good point estimation are unbiasedness, consistence, efficiency and sufficiency.	3. There is no such criteria in case of interval estimation, but, narrower the interval, better is the interval estimator at a certain confidence level.

5.7. Null hypothesis and alternative hypothesis

The hypothesis that we assume that there is no significant difference between the true value of the parameter and its hypothetical value or there is no significant difference between the true values of the two or more parameters, is called **null hypothesis**. It is denoted by H_0.

The complement of the null hypothesis, is called the **alternative hypothesis**. It is denoted by $H_1 or\ H_a$.

For instance, if θ_0 be the hypothetical value of the unknown parameter θ, then, the null hypothesis is $H_0: \theta = \theta_0$ and the alternative hypothesis is $H_1: \theta \neq \theta_0$ or $\theta > \theta_0$ or $\theta < \theta_0$. Again, if θ_1 and θ_2 be the two unknown parameters of two populations, then $H_0: \theta_1 = \theta_2$ and $H_1: \theta_1 \neq \theta_2\ or\ \theta_1 > \theta_2$ or $\theta_1 < \theta_2$.

Definition of Testing of hypothesis

It is a procedure to test whether we reject or accept the null hypothesis at a certain level of significance, based on the evidence from the sample.

5.8. Type I error and Type II error

Rejecting the null hypothesis, when it is true, is called type I error and the probability of type I error is nothing but the value of **the level of significance**. Thus,

P[type I error] = P[rejecting $\boldsymbol{H_0}$ | $\boldsymbol{H_0}$ is true] = $\boldsymbol{\alpha}$

Accepting the null hypothesis, when it is not true, is called type II error and the probability of type II error, is denoted by β. Thus,

P[type II error] = P[accepting $\boldsymbol{H_0 | H_0}$ is not true] = $\boldsymbol{\beta}$.

The complement of this probability is called the **power of the test**. That is, $\boldsymbol{1-\beta}$ is called **power of the test**. Thus, more the value of $1-\beta$ or lesser the value of β, more the power of the test.

Two tailed test and One-tailed test

If we are to test the null hypothesis $H_0: \theta = \theta_0$ against the alternative hypothesis $H_1: \theta \neq \theta_0$, then it is called **two-tailed test**.

A one tailed test may be either left-tailed test or right tailed test.

If we are to test the null hypothesis $H_0: \theta = \theta_0$ against the alternative hypothesis $H_1: \theta < \theta_0$, then it is called **left-tailed test**.

If we are to test the null hypothesis $H_0: \theta = \theta_0$ against the alternative hypothesis $H_1: \theta > \theta_0$, then it is called **right-tailed test**.

Steps / Procedures of testing of hypothesis

Step1: To set up the null and alternative hypotheses.

Step2: To find the value of the test statistic (Z or t or F or chi-square) from the sampling distribution.

Step3: To set up the level of significance α (0.01 or 0.05 or 0.10)

Step4: To find the critical value of the test statistic at α level of significance. Or to find the rejection region (p) of the test statistic.

Step5: If the value of the test statistic is less than or equal to the critical value mentioned in the step4, then we do not reject the null hypothesis. Otherwise, we reject the null hypothesis. Or if the p value mentioned in the step4 is greater than or equal to the value of the level of significance i.e., α, we do not reject the null hypothesis. Otherwise, we reject the null hypothesis.

5.9 Z-test, t-test, F-test and chi-square (χ^2)-test

Z-test and t-test

Z-test or standard normal variate test is applied when (i) the population is normal and the population standard deviation is given or (ii) the sample size 'n' is large (>30).

t-test is applied when (i) population standard deviation is not given and (ii) the sample size n is small (≤ 30).

In t-test, we apply an important term called **Degrees of Freedom (d.f.)** , which is defined as d.f. = Number of observations – number of constraints (restrictions / limitations).

In a sample of size 'n', df = n – 1, since out of the 'n' observations one observation has restriction and n – 1 observations have no restriction. In case of two independent samples of sizes n_1 and n_2, df = $n_1 + n_2 - 2$.

Test for specified Population Mean

Let μ_0 be the hypothetical value of the population mean μ. To test the null hypothesis $H_0: \mu = \mu_0$, we apply the following test statistics.

When the population is normal and the population standard deviation σ is given, then

$\boldsymbol{Z = \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}}}$ **; Z ~ N(0, 1)**

When the sample size n is large and the population standard deviation σ is not given, then

$\boldsymbol{Z = \frac{\bar{x}-\mu}{\frac{s}{\sqrt{n}}}}$ **; Z ~ N(0, 1)**

When the sample size 'n' is small (< 30) and the population standard deviation σ is not given, then

$\boldsymbol{t = \frac{\bar{x}-\mu}{\frac{s}{\sqrt{n}}}}$ **; t ~ Student's t distribution with n – 1 df.**

Where,

n = sample size

$\bar{x}$ = the sample mean = $\frac{\Sigma x}{n}$ (for ungrouped data)

= $\frac{\Sigma f x}{\Sigma f}$ (for grouped data)

s = the sample standard deviation

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

$= \sqrt{\frac{\sum f(x-\bar{x})^2}{\sum f-1}} = \sqrt{\frac{1}{\sum f-1}(\sum x^2 - \frac{(\sum fx)^2}{\sum f})}$ (for grouped data)

Ex.1. The mean life of a sample of 100 electric bulbs produced by a company is found to be 1570 hours with a standard deviation of 120 hours. Test at 5% significance level, whether the mean life of bulbs produced by the company significantly different from 1600 hours. [Given, the critical value of the test statistic at 5% significance level is 1.645 for one-tailed test and 1.96 for two-tailed test]

Sol.

STEP1. Null hypothesis H_0: the mean life of bulbs produced by the company is not significantly different from 1600 hours i.e., $\mu = 1600$

Alternative hypothesis H_1: the mean life of bulbs produced by the company is significantly different from 1600 hours i.e., $\mu \neq 1600$

STEP2. Given, the sample mean $\bar{x} = 1570$

The sample standard deviation $s = 120$

Since, the sample size n = 100, which is large (> 30), so, we apply Z-statistic.

Test statistic $Z = \frac{\bar{x}-\mu}{\frac{s}{\sqrt{n}}} = \frac{1570-1600}{\frac{120}{\sqrt{100}}} = \frac{-30}{12} = -2.5$

Therefore, $|Z| = 2.5$

STEP3. Level of significance $\alpha = 0.05$

STEP4. It is two-tailed test, since, our alternative hypothesis H$_1$: $\mu \neq 1600$. The critical value of Z is $Z_{0.05} = 1.96$

STEP5. Since, |Z| = 2.5 > $Z_{0.05} = 1.96$, so, the null hypothesis H_0 is rejected at $\alpha = 0.05$ level of significance. Hence, we conclude that the mean life of bulbs produced by the company is significantly different from 1600 hours.

Ex.2. A pharmaceutical company maintains that the mean time for a drug to take effect is 24 minutes. In a sample of 100 trials, the mean time is found to be 26 minutes with a standard deviation of 4 minutes. Can you say that the claim of the company is justified at 5% level of significance? [Given, the critical value of the test statistic at 5% significance level is 1.645 for one-tailed and 1.96 for two-tailed test]

Sol.

Null hypothesis H_0: the mean time for a drug to take effect is 24 minutes. In other words, the claim of the company is justified. i.e., $\mu = 24$

Alternative hypothesis H_1: the claim of the company is not justified. In other words, the mean time to drug to make effect is greater than 24 minutes. i.e., $\mu > 24$

Given, the sample mean $\bar{x} = 26$

The sample standard deviation $s = 4$

Since, the sample size n = 100, is large, so, we apply the test statistic Z.

The test statistic is $Z = \frac{\bar{x}-\mu}{\frac{s}{\sqrt{n}}} = \frac{26-24}{\frac{4}{\sqrt{100}}} = \frac{2}{0.4} = 5$

Level of significance $\alpha = 0.05$

It is one-tailed test, since our alternative hypothesis H_1: $\mu > 4$

The critical value of Z is $Z_{0.05} = 1.645$

Since, Z = 5 > $Z_{0.05} = 1.645$, so, the null hypothesis H_0 is rejected at $\alpha = 0.05$ level of significance. Hence, we conclude that the claim of the company is not justified. In other words, the mean time to drug to make effect is greater than 24 minutes.

Ex.3. A sample of size 10 drawn from a normal population has a mean 31 and variance 2.25. Is it reasonable to assume that the mean of the population is 30 at 1% level of significance? [Given, the critical value of t at 1% level of significance for 8 and 9 degrees of freedom are respectively 3.355 and 3.250]

Sol.

Null hypothesis H_0: It is reasonable to assume that the mean of the population is 30, i.e., $\mu = 30$

Alternative hypothesis H_1: It is not reasonable to assume that the mean of the population is 30, i.e., $\mu \neq 30$

Given, n = 10, which is small (< 30), so, we apply t-statistic

$s^2 = 2.25$ and $\bar{x} = 31$

Test statistic is $t = \frac{\bar{x}-\mu}{\frac{s}{\sqrt{n}}} = \frac{31-30}{\frac{\sqrt{2.25}}{\sqrt{10}}} = 2.11$

It is given that the level of significance $\alpha = 0.01$

Degrees of freedom (d.f.) = n – 1 = 10 – 1 = 9

It is two-tailed test, since, our alternative hypothesis H_1: $\mu \neq 30$

The critical value $t_{0.01,9} = 3.25$

Since, the calculated t = 2.11 < $t_{0.01,9} = 3.25$, so we do not reject the null hypothesis.

Hence, we conclude that it is reasonable to assume that the mean of the population is 30.

Ex.4. A random sample of 10 students had I.Q's 70, 120, 110, 101, 88, 83, 95, 98, 107 and 105. Do these data support the assumption of population mean IQ less than 100 at 5% level of significance? [Given, the critical value of t at 5% level of significance for 9 degrees of freedom is 2.262]

Sol.

H_0: the data do not support the assumption of population mean IQ less than 100. That is, $\mu = 100$

H_1: the data support the assumption of population mean IQ less than 100. That is, $\mu < 100$

$\boldsymbol{X}$	$(\boldsymbol{X} - \overline{\boldsymbol{X}})$ $= (\boldsymbol{X} - \mathbf{97.7})$	$(\boldsymbol{X} - \overline{\boldsymbol{X}})^2$
70	-27.7	767.29
120	22.3	497.29
110	12.3	151.29
101	3.3	10.89
88	-9.7	94.09
83	-14.7	216.09
95	-2.7	7.29
98	0.3	0.09
107	9.3	86.49
105	7.3	53.29
$\sum X = 977$	0	$\sum(X - \bar{X})^2$ $= 1884.1$

$\bar{X} = \frac{\sum X}{n} = \frac{977}{10} = 97.7$

Sample standard deviation

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

$$= \sqrt{\frac{1884.1}{9}} = 14.47$$

Test statistic $t = \frac{\bar{X}-\mu}{\frac{s}{\sqrt{n}}}$

$$= \frac{97.7-100}{\frac{14.47}{\sqrt{10}}}$$

$$= \frac{-2.3}{4.56} = -0.504$$

Level of significance $\alpha = 0.05$

Degrees of freedom (d.f.) = n – 1 = 10 – 1= 9

It is one-tailed test, since, our alternative hypothesis H_1: $\mu < 100$

The critical value $t_{0.05,9} = 2.262$ (one-tailed test)

∴ |t| = 0.504 < $t_{0.05,9} = 2.262$

So, we do not reject the null hypothesis.

Hence, we conclude that the data do not support the assumption of population mean IQ less than 100.

Check your progress

Ex.5. The mean breaking strength (in coded units) of the cables supplied by a manufacturer is 1800 with S.D. 100. By applying a new technique in the manufacturing process it is claimed that the breaking strength of the cables have increased. In order to test this claim a sample of 50 cables is tested. It is found that the mean breaking strength is 1850. Can we support the claim at 0.01 level of significance? [Given, the critical value of the test statistic at 0.01 level of significance is 2.33 for one-tailed test and 2.58 for two-tailed test]

Check your progress

Ex.6. A machine is designed to produce insulating washers for electrical devices of average thickness of 0.025 cm. A random sample of 10 washers was found to have an average thickness of 0.024 cm with a S.D. of 0.002 cm. Tes the significance of the deviation. [The critical value of the test statistic at 5% level of significance for 9 and 10 degrees of freedom are respectively 2.262 and 2.228]

Test for The significant difference between the Two Means

Let, μ_1 and μ_2 be the hypothetical values of two population means. To test the null hypothesis $\boldsymbol{H_0: \mu_1 = \mu_2}$ i.e., there is no significant difference between the two means, we apply the following test statistics.

When the populations are normal and the population standard deviations σ_1 and σ_2 are given, the test statistic is

$\boldsymbol{Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}}$, where, $\bar{x}_1$ and $\bar{x}_2$ are the means of the two samples of sizes n_1 and n_2 drawn from two normal populations with S.D.s σ_1 and σ_2 respectively.

If two large samples of sizes n_1 and n_2 are drawn from the two respective populations with unknown parameters, the test statistic is $\boldsymbol{Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}}$, where, s_1^2 and s_2^2 are two sample variances. s_1^2 and s_2^2 are are obtained by using the following formulae

$\boldsymbol{s_i^2} = \sqrt{\frac{\sum(x_i - \bar{x}_i)^2}{n_i - 1}} = \sqrt{\frac{1}{n_i - 1}\left(\sum \boldsymbol{x_i^2} - \frac{(\sum x_i)^2}{n_i}\right)}$ (for ungrouped data)

$\boldsymbol{s_i^2} = \sqrt{\frac{\sum f_i(x_i - \bar{x}_i)^2}{\sum f_i - 1}} = \sqrt{\frac{1}{\sum f_i - 1}\left(\sum \boldsymbol{x_i^2} - \frac{(\sum f_i x_i)^2}{\sum f_i}\right)}$ (for grouped data)

The statistic Z ~ N(0, 1)

When the samples are small and the population standard deviations σ_1 and σ_2 are not given, i.e., unknown, two small samples of sizes n_1 (< 30) and n_2 (< 30) are drawn from two normal populations with unknown standard deviations σ_1 and σ_2, we apply the following t-statistic.

$\boldsymbol{t} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{S^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$, where, S^2 is the pooled sample variance, which is obtained by using the following formula

$\boldsymbol{S^2} = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$

It is assumed that $\sigma_1^2 = \sigma_2^2$

The statistic t follows Student's t distribution with $\boldsymbol{n_1 + n_2 - 2}$ degrees of freedom (df).

Ex.7. The mean of two samples of sizes 150 and 200 are respectively 67.5 and 68. Their respective standard deviations are 3 and 2.5. Is there significant difference between the two means at 5% level of significance. [Critical value of the test statistics at 5% level of significance, are 1.645 for one-tailed test and 1.96 for two-tailed test]

Sol.

Null hypothesis H_0: there is no significant difference between the means, i.e., $\mu_1 = \mu_2$

Alternative hypothesis H_1: there is significant difference between the means, i.e., $\mu_1 \neq \mu_2$

Given, the sample sizes $n_1 = 150, n_2 = 200$

Since, the samples are large (both the sample sizes > 30), so, we use Z-test.

We have, the sample means $\bar{x}_1 = 67.5, \bar{x}_2 = 68$

Sample standard deviations $s_1 = 3, s_2 = 2.5$

Test statistic $Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{67.5-68}{\sqrt{\frac{3^2}{150}+\frac{2.5^2}{200}}} = -1.655$

|Z| = 1.655

Level of significance $\alpha = 0.05$

It is two-tailed test, since, our alternative hypothesis is $\mu_1 \neq \mu_2$

So, the critical value of Z at $\alpha = 0.05$ for two-tailed test is $|Z_{0.05}| = 1.96$

∴ The test statistic value |Z| = 1.655 < 1.96

Therefore, we do not reject the null hypothesis.

Hence, we conclude that there is no significant difference between the means.

Ex.8. It is desired to test if there is any significant difference between the average ages of female and male students of a particular class in an educational institution. A random sample of 10 female students reveals that the average age is 23 years with a standard deviation 4 years. A random sample of 8 male students reveals an average of 26 years with a standard deviation of 5 years. Test at 5% level whether there is any significant difference between the average age of female and male students. [The critical values of the test statistic at $\alpha = 0.05$ for 16, 17 and 18 degrees of freedom are respectively 2.120, 2.110 and 2.101]

Sol.

Null hypothesis H_0: There is no significant difference between the average age of female and male students, i.e., $\mu_1 = \mu_2$.

Alternative hypothesis H_1: There is significant difference between the average age of female and male students, i.e., $\mu_1 \neq \mu_2$.

Given,	Female students	Male students
Sample size	$n_1 = 10$	$n_2 = 8$
Sample mean (years)	$\bar{x}_1 = 23$	$\bar{x}_2 = 26$
Sample SD	$s_1 = 4$	$s_2 = 5$

Since, both the sample sizes are small (< 30), so, we apply t-statistic.

Pooled sample variance $s^2 = \frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}$

$$= \frac{(10-1)\times 4^2+(8-1)\times 5^2}{10+8-2}$$

$$= \frac{319}{16} = 19.9375$$

Test statistic is $t = \frac{\bar{x}_1-\bar{x}_2}{\sqrt{s^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$

$$= \frac{23-26}{\sqrt{19.9375\times\left(\frac{1}{10}+\frac{1}{8}\right)}}$$

$$= -\frac{3}{\sqrt{4.4859375}}$$

$$= -\frac{3}{2.1180} = -1.42$$

Degrees of freedom (df) = $n_1 + n_2 - 2$

= 10 + 8 – 2 = 16

Level of significance $\alpha = 0.05$

It is two-tailed test, since, our alternative hypothesis $\mu_1 \neq \mu_2$

∴ The critical value oof t at $\alpha = 0.05$ for 16 df in two-tailed test is 2.120

Therefore, |t| = 1.42 < 2.120. so, we do not reject the null hypothesis H_0.

Hence, we may conclude that there is no significant difference between the average age of female and male students.

Check your progress

Ex.9. Electric bulbs manufactured by X and Y companies gave the following results.

	X	**Y**
Number of bulbs used	**100**	**100**

Mean life in hours	**1300**	**1248**
S.D. in hours	**82**	**93**

Test whether there is any significant difference in the mean life of two makes at 5% level of significance? [Given, the critical values of the test statistic at $\alpha = 0.05$ in one-tailed and two-tailed tests are respectively 1.645 and 1.96]

Check your progress

Ex.10. Two types of batteries are tested for their length of life and the following data are obtained.

	Sample size	Mean life	Variaance
Type A	**9**	**600**	**121**
Type B	**8**	**640**	**144**

Is there a significant difference in the two means? [Given, the critical values of the test statistic at 5% level of significance for 15, 16 and 17 df are respectively 2.131, 2.130, 2.110]

Paired t-test for the difference of means.

Let us consider pairs of observations (x$_i$, x$_j$) which are not independent, obtained from a sample of size ‘n’. If we are to test the null hypothesis H_0: the two means are not differ significantly, i.e., $\mu_d = 0$, we use the following test statistic

$t = \frac{\bar{d}}{\sqrt{\frac{s_d^2}{n}}}$, which follows Student’s t-distribution with n – 1 degrees of freedom (df), where, d = Difference between the pair of observation = x$_i$ - x$_j$

$\boldsymbol{\bar{d} = \frac{\sum d}{n}}$, the mean of d

$\boldsymbol{s_d^2 = \frac{\sum(d-\bar{d})^2}{n-1} = \frac{1}{n-1}\left(\sum d^2 - \frac{(\sum d)^2}{n}\right)}$, the variance of d

Ex.11. The sales data of an item in shops before and after a special promotional campaign are

Shops	**: A**	**B**	**C**	**D**	**E**	**F**
Before the campaign:	**53**	**28**	**31**	**48**	**50**	**42**
After the campaign :	**58**	**29**	**30**	**55**	**56**	**45**

Can the campaign be judged to be a success? Test at 5% level of significance. [Given, the critical value of t at $\alpha = 0.05$ for 5 df in one-tailed test and two-tailed test are respectively

2.015 and 2.571]

Sol.

Null hypothesis H_0: The campaign cannot be judged to be a success. In other words, the sales of the item remain same before and after the campaign, i.e., $\mu_d = 0$

Alternative hypothesis H_1: The campaign can be judged to be a success. In other words, the sales of the item increases after the campaign, i.e., $\mu_d < 0$.

Shops	Sales		d	d^2
	Before the campaign	After the campaign		
A	53	58	-5	25
B	28	29	-1	1
C	31	30	1	1
D	48	55	-7	49
E	50	56	-6	36
F	42	45	-3	9
Total			-21	121

Thus, we have

n = 6, $\sum d = -21$ and $\sum d^2 = 121$

$\therefore \bar{d} = \frac{\sum d}{n} = -\frac{21}{6} = -3.5$

And $s_d^2 = \frac{1}{n-1}\left(\sum d^2 - \frac{(\sum d)^2}{n}\right)$

$= \frac{1}{6-1} \times \left(121 - \frac{(-21)^2}{6}\right)$

$= \frac{1}{5} \times 47.5$ = 9.5

Test statistic is

$$t = \frac{\bar{d}}{\sqrt{\frac{s_d^2}{n}}} = -\frac{3.5}{\sqrt{\frac{9.5}{6}}} = -\frac{3.5}{1.2583} = -2.78$$

∴ |t| = 2.78

df = n – 1 = 6 – 1 = 5

Level of significance $\alpha = 0.05$

It is two-tailed test, since, our alternative hypothesis $H_1: \mu_d < 0$.

∴ The critical value of t at $\alpha = 0.05$ for 5 df in one-tailed test is 2.015.

So, |t| = 2.78 > 2.015

∴ We reject the null hypothesis H_0 at $\alpha = 0.05$.

Hence, we conclude that the campaign can be judged to be a success.

Check your progress

Ex.12. XYZ physical fitness Centre claims that competition of their weight loss programme will result in a loss of weight. To test this claim, six persons were selected at random and they were put through the weight loss programme and their weights before and after the programme were recorded. The weights in pounds of these six persons recorded before and after the programme are as follows

Person :	**1**	**2**	**3**	**4**	**5**	**6**
Weight (before):	**145**	**200**	**160**	**185**	**164**	**175**
Weight (After) :	**143**	**190**	**165**	**183**	**160**	**176**

Test the claim of the fitness Centre at 5% level of significance. [Given, the critical value of the test statistic at 5% level of significance for 5 df in one-tailed test and two-tailed test are respectively 2.015 and 2.571].

Chi-square (χ^2) test

It is a test of the difference between actual (true) observations and expected (hypothetical observations.It is a general test, since it is applied in both parametric and non-parametric cases. In chi-square test, the null hypothesis is H_0: there is no significant difference between the true and expected observations. Let $O_1, O_2, \ldots O_n$ be the true (or actual) observations of an experiment and $E_1, E_2, \ldots E_N$ be their corresponding expected (or hypothetical) observations so that $\sum_{i=1}^{n} O_i = \sum_{i=1}^{n} E_i$. Then the test statistic is $\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$, which follows a chi-square distribution with **n – 1 degree of freedom**. If the value of the test statistic ≤ the critical value of the chi-square at α leve of significance with n – 1 degree of freedom ($\chi^2_{\alpha,n-1}$), we do not reject the null hypothesis, otherwise, we reject the null hypothesis at αlevel of significance.

Chi-Square (χ^2) Test for the "Goodness of fit"

In this case, we have to test the hypothesis H_0: the fit is good i.e, the observations follow a certain distribution. Then the test statistic is $\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$, which follows a chi-square distribution with n – 1 degrees of freedom.

Ex.13. The following table gives the number of aircraft accidents that occurred during the six days in a week. Find at 5% level of significance, whether the accidents are uniformly distributed over the week.

Days	**:Mon**	**Tue**	**Wed**	**Thu**	**Fri**	**Sat**
No. of accidents:	**14**	**18**	**12**	**11**	**15**	**14**

[Given, the critical values of χ^2 at 5% level of significance for 5 and 6 degrees of freedom are respectively 11.07 and 12.59]

Sol.

Null hypothesis H_0: the accidents are uniformly distributed.

Alternative hypothesis H_1: the accidents are not uniformly distributed.

Calculations of chi-square (χ^2)

DAYS	**NO. OF ACCIDENTS (O_i)**	**E_i**	**(O_i – E_i)**	$\frac{(O_i - E_i)^2}{E_i}$
MON	14	$\frac{84}{6} = 14$	0	0
TUE	18	$\frac{84}{6} = 14$	4	$\frac{16}{14}$
WED	12	$\frac{84}{6} = 14$	-2	$\frac{4}{14}$
THU	11	$\frac{84}{6} = 14$	-3	$\frac{9}{14}$
FRI	15	$\frac{84}{6} = 14$	1	$\frac{1}{14}$

SAT	14	$\frac{84}{6}$ $= 14$	0	0
TOTAL	$\sum \boldsymbol{O_i = 84}$	$\sum \boldsymbol{E_i}$ $\boldsymbol{= 84}$	**0**	$\sum \frac{\boldsymbol{(O_i - E_i)^2}}{\boldsymbol{E_i}}$ $\boldsymbol{= \frac{30}{14} = 2.143}$

Test statistic $\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} = \frac{30}{14} = 2.143$

No. of observations n = 6

So, degree of freedom (df) = n – 1 = 6 – 1 = 5

Given, the critical value of χ^2 at 5% level of significance and 5 degree of freedom is 11.07 i.e, $\chi^2_{0.05,5} = 11.07$

The value of the test statistic $\chi^2 = 2.143 < \chi^2_{0.05,5} = 11.07$

Therefore, the null hypothesis H_0 is not rejected at α = 0.05 level of significance.

Hence, we conclude that the accidents are uniformly distributed.

Check your progress

Ex.14. A dice is thrown 60 times and the following results are obtained.

Face :	**1**	**2**	**3**	**4**	**5**	**6**
Frequency:	**8**	**7**	**12**	**8**	**14**	**11**

Test at 5% level of significance, if the dice is unbiased.

[Given, $\boldsymbol{\chi^2_{0.05,5} = 11.07}$]

Chi-Square (χ^2) Test for the "Independence of attributes"

In this case, the entire observations are divided according the the attributes in **r ×c contingency table** as shown in the figure, where, r = the number of rows and c = the number of columns. Let O$_{ij}$ be the true obsevations of the ith row and jth column, E$_{ij}$, the corresponing expected observations, where, i = 1, 2, 3,…,r and j = 1, 2, 3, …, c. It is assumed that $\sum O_{ij} = \sum E_{ij} = N$

					Total
	O$_{11}$ E$_{11}$	O$_{12}$ E$_{12}$	…..	O$_{1c}$ E$_{1c}$	R$_1$

	O_{21} E_{21}	O_{22} E_{22}	….	O_{2c} E_{2c}	R_2
	. . .	. . .	….. ….. …..	. . .	. . .
	O_{r1} E_{r1}	O_{r2} E_{r2}	…..	O_{rc} E_{rc}	R_r
Total	C_1	C_2	…..	C_c	N

The expected frequency Eij is obtained by using the following method

$E_{ij} = \frac{R_i \times C_j}{N}$, where, R_i = the ith trow total, C_j = jth column total

We have to test the null hypothesis H_0: the attributes are independent. To test this null hypothesis, the following test statistic is used.

$\chi^2 = \sum\sum \frac{(O_{ij}-E_{ij})^2}{E_{ij}}$, which follows a chi-square distribution with (r – 1) ×(c – 1) degree of freedom.

Ex.15. To test the efficiency of a new drug a controlled experiment was conducted wherein 300 patients were administered the new drug and 200 other patients were not given the drug. The patients were monitored and results were obtained as follows:

	Cured	**Condition worsened**	**No effect**
Given the drug	**200**	**40**	**60**
Not given the drug	**120**	**30**	**50**

[Given,	**$\chi^2_{0.05}$ =**	**3.84**	**5.99**	**7.8**
	d.f. =	**1**	**2**	**3]**

Sol.

Null hypothesis H_0: The new drug is not efficient. In other words, the drug is not effective.

Alternative hypothesis H_1: The new drug is efficient. In other words, the drug is effective.

Given, two way contingency table

	Cured	**Condition worsened**	**No effect**	**Total**
Given the drug	200	40	60	**300**
Not given the drug	120	30	50	**200**
Total	**320**	**70**	**110**	**500**

Calculations of test statistic

Observed frequency O_{ij}	Expected frequency E_{ij}	O_{ij} - E_{ij}	$\frac{\left(O_{ij}-E_{ij}\right)^2}{E_{ij}}$
O_{11} = 200	$E_{11}=\frac{300\times 320}{500}=192$	8	0.333
O_{12} = 40	$E_{12}=\frac{300\times 70}{500}=42$	-2	0.095
O_{13} = 60	$E_{13}=\frac{300\times 110}{500}=66$	-6	0.545
O_{21} = 120	$E_{21}=\frac{200\times 320}{500}=128$	-8	0.500
O_{22} = 30	$E_{22}=\frac{200\times 70}{500}=28$	2	0.143
O_{23} = 50	$E_{23}=\frac{200\times 110}{500}=44$	6	0.818
$\sum O_{ij}=500$	$\sum E_{ij}=500$	0	$\sum\frac{\left(O_{ij}-E_{ij}\right)^2}{E_{ij}}$ $=2.434$

Therefore, calculated $\chi^2=\sum\frac{\left(O_{ij}-E_{ij}\right)^2}{E_{ij}}=2.434$

Degree of freedom (d.f.) = (r – 1)×(c – 1)

= (2 – 1)×(3 – 1) = 2

[since, no. of rows r = 2, no. of columns c = 3 (only figures, excluding total)]

Level of significance $\alpha=0.05$

The critical value is ${\chi^2}_{0.05}$= 5.99

So, $\chi^2=2.434$ < 5.99

Therefore, the null hypothesis H_0 is not rejected at $\alpha=0.05$

Hence, we may conclude that the new drug is not efficient. In other words, the drug is not effective.

Check your progress

Ex.16. A movie producer is bringing out a new movie. In order to map out is advertising campaign he wants to determine whether the movie will appeal most to a particular age group or whether it will appeal equal to all age groups. The producer takes a random sample from persons attending a preview of the movies and obtains the following results. Use test at 5% level to derive the conclusion.

	Age group			
	Under 20	**20 - 39**	**40 - 59**	**60 and above**

Liked the movie	**320**	**80**	**110**	**200**
Disliked the movie	**50**	**15**	**70**	**60**
Indifferent	**30**	**5**	**20**	**40**

[Given that $\chi^2_{0.05,6} = 12.59$]

Uses of chi-square (χ^2) test

The χ^2 test is very powerful test as it is partially used in case of parametric distribution and non-parametric distribution. It is used to estimate the population variance, not the population mean or proportion. The χ^2test is independent of the population distribution. The important uses of χ^2test are

(i) To test the discrepancy between observed and expected frequencies. χ^2 test determines the degree of deviation between observed frequencies and the theoretical frequencies and to conclude whether the deviation is due to chance or not.

(ii) To determine the association between two or more attributes. By using the χ^2test, we can find out whether two or more attributes are associated or not.

F-test or Variance Ratio test

F-test or Fisher's test is used to compare two variances. If we are to test the null hypothesis H$_0$: the two variances are equal i.e, $\sigma_1^2 = \sigma_2^2$ against the alternative hypothesis H$_1$: $\sigma_1^2 \neq \sigma_2^2$(two tailed) or $\sigma_1^2 > \sigma_2^2$(right tailed test) or $\sigma_1^2 < \sigma_2^2$(left tailed), we use the following test statistic $\boldsymbol{F = \frac{S_1^2}{S_2^2}, if\ S_1^2 > S_2^2}$. The statistic F follows F-distribution with $(n_1 - 1, n_2 - 1)$ degree of freedom, where, S_1^2 and S_2^2 are the variances of the samples of sizes n_1 and n_2 respectively.

OR

$\boldsymbol{F = \frac{S_2^2}{S_1^2}, if\ S_1^2 < S_2^2}$. F follows F-distribution with $(n_2 - 1, n_1 - 1)$ degree of freedom.

If the value of the test statistic F is less than or equal to its critical value at α level of significance, we do not reject the null hypothesis H$_0$, otherwise we reject the null hypothesis H$_0$.

F-test is also used to test the equality of several means of the populations (at least two population means) i.e, to test the null hypothesis H$_0$: $\mu_1 = \mu_2 = \cdots = \mu_n$ against the alternative hypothesis H$_1$: at least two means are not equal. This test is done by using the analysis of variance (ANOVA) with the help of the test statistic $\boldsymbol{F} = \frac{\boldsymbol{variance\ between\ the\ samples}}{\boldsymbol{vaariance\ within\ the\ samples}}$.

Assumptions of F-test

(i) The samples are independent and are drawn from two normal populations

(ii)The variances of the two populations are assumed to be equal, i.e., $\sigma_1^2 = \sigma_2^2$

(iii) The ratio between the two variances should be ≥ 1

Ex.17. Two samples of sizes 10 and 15, are drawn from two populations of unknown variances. The variances of the two samples are 100 and 144. Test at 5% level of significance, whether the two variances are equal or not. [Given, the critical value of F at 5% level of significance with (9, 14) degree of freedom is 2.59 and with (14,9) degree of freedom is 3.04]

Sol.

Null hypothesis H_0: the variances are equal i.e, $\sigma_1^2 = \sigma_2^2$

Alternative hypothesis H_1: the variances are not equal i.e, $\sigma_1^2 \neq \sigma_2^2$

Given, the sample sizes $n_1 = 10, n_2 = 15$

Sample variances $S_1^2 = 100,\ S_2^2 = 144$

Since, $S_1^2 < S_2^2$, so the test statistic is $F = \frac{S_2^2}{S_1^2} = \frac{144}{100} = 1.44$.

Degree of freedom $= (n_2 - 1, n_1 - 1) = (15 - 1, 10 - 1) = (14,9)$

Level of significance α = 5% = 0.05

So, the critical value of the test statistic F at 5% level of significance with (14,9) degree of freedom is $F_{0.05}(14,9) = 3.04$.

Since, the value of the test statistic F = 1.44 < 3.04

So, we do not reject the null hypothesis H0 at 5% level of significance.

Hence, we conclude that the variances are equal.

Check your progress

Ex.18. The following figures relate to the number of units of an item produced per shift by two workers A and B for a number of days.

A: 16 17 18 19 20 21 22 24 26 29

B: 19 22 23 25 26 28 29 30 31 32 35 36

Can it be inferred that worker A is more stable compared to worker B? Give your answer

using F-test at 5% level of significance. [Given, $\boldsymbol{F_{0.05}(11, 9) = 3.16}$]

Hints: H_0: The stabikity of both the workers A and B are equal, i.e., $\sigma_1^2 = \sigma_2^2$, H_1: Worker A is more stable compared to worker B, i.e., $\sigma_1^2 < \sigma_2^2$. The sample variance of the worker A is $s_1^2 = \frac{1}{(n_1-1)}\sum(x_1 - \bar{x}_1)^2 = 14$ and the sample variance of the worker B is $s_2^2 = \frac{1}{n_2-1}\sum(x_2 - \bar{x}_2)^2 = 27.09$. Then, F = $\frac{s_2^2}{s_1^2}$, since, $s_2^2 > s_1^2$ with df =(n$_2$-1, n$_1$-1)]

-----XXXXX-----

Techno City, Khanapara, Kling Road, Baridua, 9th Mile,
Ri-Bhoi, Meghalaya-793101
Phone: 9508 444 000, Web : www.ustm.ac.in