14.5.	**Statistical Package for the Social Sciences (SPSS)**	287
14.5.1.	Introduction	287
14.5.2.	Starting SPSS	287
14.5.3.	Creating and Entering Data in SPSS	289
14.5.4.	Computation of Descriptive Statistics	295
14.6.	**Minitab**	303
14.6.1.	Introduction	303
14.6.2.	Entering Data	303
14.6.3.	Viewing Descriptive Statistics	305
14.6.4.	Creating Graphs and Charts	307
14.6.5.	Running A Regression Analysis	310
14.7.	**Design of Experiment**	312
14.7.1.	Introduction	312
14.7.2.	Components of Experimental Design	312
14.7.3.	Purpose of Experimentation	313
14.7.4.	Experiment Design Process	313
14.8.	**Online Statistical Software's to Industrial and Clinical Trial Approach**	314
14.9.	**Exercise**	315

Module 5

Chapter 15: Design and Analysis of Experiemnt

15.1.	**Factorial Design**	315
15.1.1.	Introduction	315
15.1.2.	Definition (Vocabulary)	315
15.1.3.	2^2 Factorial Design	317
15.1.4.	2^3 Factorial Design	320
15.1.5.	Advantages of Factorial Design	321
15.1.6.	Disadvantages of Factorial Design	322
15.2.	**Response Surface Methodology**	322
15.2.1.	Introduction	322
15.2.2.	Objectives of Response Surface Methodology (RSM)	323
15.2.3.	Experimental Strategy	323
15.2.4.	Types of Models	324
15.2.5.	Sequential Nature of RSM	324
15.2.6.	Methods of RSM	324
15.2.7.	Applications of RSM	324
15.3.	**Response Surface Designs**	325
15.3.1.	Introduction	325
15.3.2.	Types of Response Surface Design	325
15.4.	**Historical Data Design**	329
15.4.1.	Introduction	329
15.4.2.	Design the "Experiment"	330
15.4.3.	A Peculiarity on Pasting Data	331
15.4.4.	Analyze the Results	331
15.5.	**Optimization Techniques**	337
15.5.1.	Introduction	337
15.5.2.	Optimization Using Factorial Designs	339
15.5.3.	Sequential Optimization	343
15.5.4.	Screen Design	347
15.6.	**Exercise**	349

CHAPTER 1 Introduction to Biostatistics

1.1. STATISTICS AND BIOSTATISTICS

1.1.1. Meaning and Definition

The origin of the word 'statistics' can be traced back to either the Latin word *Status* or the Italian word *Statista* or German word *Statistik*. All the three words mean, 'an organised political state'. 'Statistics' was regarded as the 'science of statecraft' in early years because it was mainly used by the administration to keep track of births, death, crop yield, taxes, etc. It was referred to as the 'science of king' or 'king' as it was used as a tool for government record collection. In present time, statistics is regarded as an important tool for decision-making in uncertain times.

According to the first concept (Descriptive Statistics or Statistical data), statistics is expressed as pertaining to numerical data. This concept takes statistics in the plural sense.

According to the second concept (Statistical methods) statistics is expressed as a science. It takes 'Statistics' in the singular sense.

1.1.1.1. Statistics as Statistical Data

Statistical data refers to the collection of quantitative data such as price of a commodity, production, export, import, births, deaths, etc.

According to Bowley, "Statistics are numerical statement of facts in any department of enquiry placed in relation to each other". This definition gives importance to quantitative aspects and helps in comparative study of figures.

According to Webster, "Classified facts respecting the conditions of the people in a state especially those facts which can be stated in numbers or in any other tabular or classified arrangements". The above definition denotes that the facts which are expressed quantitatively can be termed as statistics and qualitative facts cannot be treated statistically.

According to Yule and Kendall, "By statistics we mean quantitative data affected to a marked extent by multiplicity of cause".

According to Boddington, "Statistics is a science of estimates and probabilities".

For the analysis of biological phenomena, biostatistics (or biometry) deals with the statistical process and methods applied. Biostatistics includes the design of biological experiments and demonstrates the collection summarization, and analysis of data from those experiments.

1.1.1.2. Statistics as Statistical Methods

Statistics as a 'science of statistical methods' is based on the idea that, "Statistics is what statistics does or statistics is what statisticians do". This definition considers statistics as a subject or method.

According to A.L. Bowley, "Statistics may be called the science of counting". In this definition only one aspect is covered, i.e., counting. But, in many cases data collection also depends on estimation, which helps in further classification, tabulation, etc.

Scanned with CamScanner

According to A.L.Bowley, "Statistics may rightly be called the science of averages". Statistics uses averages to summarise the collected data. However, it also uses diagrams, graphs, correlation coefficient, etc. for analysis, which has not been included in the above definition of statistics and hence make it incomplete.

According to Wallis and Roberts, "Statistics is a body of methods for making decisions in the face of uncertainty".

According to Croxton and Cowden, "Statistics may be defined as the collection, presentation, analysis and interpretation of numerical data".

1.1.2. Types of Statistics

Statistics can be classified into two main categories:

1) **Descriptive Statistics:** This includes the methods of data collection, presentation and characterisation of a set of data. All these help in describing the various features of the collected sample data. Descriptive statistics includes graphical representation and quantitative measures. **For example,** graphical representation includes bar charts, line graphs and pie charts whereas measures of central tendency, dispersion, skewness and kurtosis denote quantitative measures.
2) **Inferential Statistics:** Inferential statistics includes the methods that help in characterising a population or help in decision-making which is based on the sample results of the population. Sample and population are two relative terms. The larger unit about which analysis is to be done is referred to as 'population' or 'universe' and the fraction or a portion of that population/universe is called 'sample'.

1.1.3. Characteristics of Statistics

Basic characteristics of statistics are as follows:

1) Statistics are averages of numerical data. Single or isolated figures cannot be considered as statistics as they cannot help in analysis or drawing of conclusions. Single death or birth does not comprise statistics but a number of such figures help in determining the average death or birth rate.
2) Generally, statistics contains quantitative information which means it is always expressed numerically. Qualitative facts can never be measured numerically and therefore it cannot give accurate interpretations. Hence, this can never be considered as statistics.
3) Statistics are affected by various causes or factors. One cause or factor cannot be said to be responsible for the results after analysis of a given set of data.
4) Statistical data is collected in a systematic manner for a predetermined purpose. This means that the reason and plan for data collection is predetermined.
5) Statistical data may be used for comparative studies. **For example,** for comparing the efficiency of two departments in an office, the performance appraisal can be used.

1.1.4. Scope/Applications of Statistics

Statistics has affected almost all areas of life as it covers simple households to big businesses, even government also. Some of the areas where statistics has been used are as follows:

1) **Statistics and the Government:** Statistics has been used extensively since the beginning of organised society. Statistics has been used by the administrative heads and rulers of the states in the form of collecting data on different aspects for the purpose of formulating sound military and fiscal policies. This data includes figures of population, tax collection, military strength, etc. In the present times, the government is the biggest collector of data as well as the biggest user of statistics. A huge amount of data is collected and interpreted by various departments of the [illegible] efficient policies and decision-making.

2) **Statistics and Mathematics:** Statistics can be considered as a branch of science which is conceived on the foundation of mathematics. A person should have some knowledge of mathematics for understanding the basic fundamentals of statistics.

   According to Connor, "Statistics is a branch of Applied Mathematics which specialises in data".

   According to W. I. King, "Statistics may properly be considered as a branch of mathematics in as much as it attempts to formulate definite rules of procedure applicable in handling groups of data of many different varieties".
3) **Statistics and Economics:** Statistics is used as an important tool in economics study and research. Economics is mainly concerned with production and distribution of wealth and also savings and investments. Statistical tools are used in the following economic interest areas:
   i) Statistical methods are used for measuring and forecasting Gross National Product (GNP).
   ii) Statistical studies of business cycles give a clear picture about the economic stability.
   iii) Economic policy-making depends on the statistical analysis of population growth, unemployment figures, rural or urban population shifts, etc.
   iv) Optimum utilisation of resources is possible with the use of econometric models which uses statistical methods.
   v) For the study of finance, banking, consumer savings and credit availability, financial statistics is necessary.
4) **Statistics and Physical Science:** The use of statistical methods is continuously increasing in the field of physical sciences such as Biology, Physics, Chemistry, Astronomy, Medicine, etc. Statistical data are collected from different results of different experiments. These are then comparatively studied to find out whether the differing results are statistically significant or not. As more facts are discovered, the different estimates in physical science such as distance between two planets in the universe become more accurate.
5) **Statistics and Natural Sciences:** Statistics is also very important in the study of natural sciences such as astronomy, biology, medicine, meteorology, zoology, botany, etc. **For example,** for diagnosing the exact disease of a patient, the doctor must believe on real data such as the body temperature, pulse rate, blood pressure, etc.
6) **Statistics and Research:** In the current scenario statistics is an essential part of research study. Improvement in knowledge has been possible because experiments are carried out with the help of statistical methods. **For example,** experiments about crop yields and their correlation with different types of fertilisers and different types of soil are designed and studied with the help of statistical techniques. In the current time, statistical methods are used in all types of research work including medicine and public health.
7) **Statistics in Astronomy:** Astronomers were one of the first groups of people who used statistics in the study of movement of heavenly bodies and eclipses and other such astronomical issues. Astronomers earlier relied heavily on estimation but later statistics helped to turn these estimations into accurate ideas.
8) **Statistics in Education:** Statistics is used extensively in the field of education because research has become a common feature in all branches of activities. In education, statistics plays a vital role in formulation of new policies for new courses to be started, consideration of infrastructural requirements for new courses, etc. Apart from this, there are many people involved in research work who test past knowledge and develop new knowledge with the help of statistics.

Scanned with CamScanner

9) **Statistics in Accounting and Auditing:** In accounting, exactness is an essential component but for decision-making purposes, approximation is taken into account. The current asset value is calculated on the basis of its current values and the corrected values are determined with the help of current purchasing power of money or the current value of it, while taking depreciation into consideration. This is done through the use of price indices which are based on the collection of statistics. For determining the trend of future profits it is required to use the study of correlation analysis between the profits and dividends. In auditing, sampling is generally used as it is not possible to examine voluminous transactions due to lack of human resources. An auditor will first co-relate the past error percentage with the current error rate after conducting a pilot audit. After this, the auditor decides on the sample size of books to be audited.

10) **Statistics in Planning:** Statistics is important for efficient planning in all modern economies, especially in the developing countries. This helps in successful planning by taking into consideration the correct analysis of complex statistical data. The plans, adopted for the economic development of a country, are also made on the basis of statistics available about the different economic problems being faced.

11) **Statistics and Commerce:** Commerce is now very much dependent on statistics for success of businesses. Any businessman cannot afford to over-stock or under-stock the goods. In the initial stage it is required to do a market survey using statistical techniques to understand the changing tastes of the consumers because a number of multinational companies have also entered with new products and services. Thus, the statistical techniques help in providing the present conditions and forecast the future.

12) **Other Areas:** Statistics is used by many businesses such as insurance companies, stock brokerage houses, banks, public utility companies and so on. It also helps politicians to study their winning chances from a constituency through the use of sampling techniques in random selection of voter samples and studying their preferences on issues and policies.

1.1.5. Uses of Biostatistics

1) Normalcy is defined by biostatistics.
2) Biostatistics is used to check whether the difference between two populations is real or a chance occurrence for a particular attribute.
3) It studies the correlation between two or more attributes in the same population.
4) By control studies, it evaluate the efficiency of vaccines, sera etc.
5) It is used to locate, define and measure the extent of morbidity and mortality in the community.
6) It is also used to evaluate the achievements of public health programs.
7) It is used to fix priorities in public health programs.

1.2. FREQUENCY

1.2.1. Introduction

Frequency is the number of occurrences of a repeating event per unit time. **For example**

1) The frequency distribution of income in a population would show how many individuals (or households) have the income of a certain level (say, 5,000 a month).
2) Consider the following data:
   Scores: 1, 1, 2, 2, 2, 2, 2, 3, 3, 3, 3, 4, 4, 5
   Frequency of 2 is 5

1.2.2. Frequency Distribution

Frequency distribution is a series when a number of observations with similar or closely related values are put in separate bunches or groups, each group being in order of magnitude in a series. It is simply a table in which the data are grouped into classes and the numbers of cases which fall in each class are recorded. It shows the frequency of occurrence of different values of a single phenomenon.

According to Errieker, 'a classification according to the number possessing same value of the variable'.

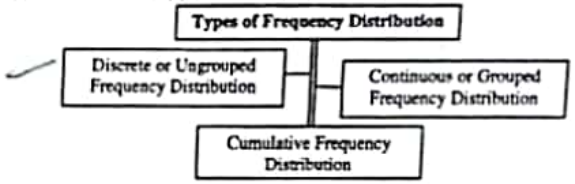
According to Croxton and Cowden, "Frequency distribution is a statistical table which shows the set of all distinct values of the variable arranged in order of magnitude, either individually or in groups, with their corresponding frequencies side by side".

Reasons for Constructing Frequency Distribution

1) To facilitate the analysis of data.
2) To estimate frequencies of the unknown population distribution from the distribution of sample data.
3) To facilitate the computation of various statistical measures.

1.2.3. Types of Frequency Distribution

Basically there are three type of frequency distribution:



1.2.3.1. Discrete or Ungrouped Frequency Distribution

In this form of distribution, the frequency refers to discrete value. Here the data are presented in a way that exact measurement of units is clearly indicated. There is definite difference between the variables of different groups of items. Each class is distinct and separate from the other class. Non-continuity from one class to another class exists. Data such as facts like the number of rooms in a house, the number of companies registered in a country, the number of children in a family, etc.

The process of preparing this type of distribution is very simple. We have just to count the number of times a particular value is repeated, which is called the frequency of that class. In order to facilitate counting prepare a column for tally marks. In another column, place all possible values of variable from the lowest to the highest.

Then put a bar (Vertical line) opposite the particular value to which it relates. To facilitate counting, blocks of five bars are prepared and some space is left in between each block. We finally count the number of bars and get frequency.

Example 1: In a survey of 40 families in a village, the number of children per family was recorded and the following data obtained.

1	0	3	2	1	5	6	2
2	1	0	3	4	2	1	6
3	2	1	5	3	3	2	4
2	2	3	0	2	1	4	5
3	3	4	4	1	2	4	5

Represent the data in the form of a discrete frequency distribution.

Scanned with CamScanner

Solution: Frequency distribution of the number of children

Number of Children	Tally Marks	Frequency
0	𝍸	3
1	𝍸 ||	7
2	𝍸 𝍸	0
3	𝍸 |||	8
4	𝍸 |	6
5	||||	4
6	||	2
Total		40

1.2.3.2. Continuous or Grouped Frequency Distribution

Continuous series is one where measurements are only approximations and are expressed in class intervals, i.e., within certain limits. **According to Boddington,** "the variable which can take any intermediate value between the smallest and longest value in the distribution".

In a continuous frequency distribution the class intervals theoretically continue from the beginning of the frequency distribution to the end without break.

Difference between Ungrouped and Grouped Frequency Distribution

The continuous frequency distribution can always be **distinguished** from the discrete frequency distribution in that it will contain two limits upper limits and lower limits of each class interval while the discrete frequency distribution will possess only one list of classification of values.

For example, Wage distribution of 100 employees

Weekly Wages (₹)	Number of Employees
50-100	4
100-150	12
150-200	22
200-250	33
250-300	16
300-350	8
350-400	5
Total	100

Basic Component of Continuous Frequency Distribution

The following are some basic technical terms when a continuous frequency distribution is formed or data are classified according to class intervals.

1) **Class Limits:** The class limits are the lowest and the highest values that can be included in the class. **For example,** take the class 30-40. The lowest value of the class is 30 and highest class is 40. The two boundaries of class are known as the **lower limits** and the **upper limit** of the class. The lower limit of a class is the value below which there can be no item in the class. The upper limit of a class is the value above which there can be no item to that class. The way in which class limits are stated depends upon the nature of the data. In statistical calculations, lower class limit is denoted by **L** and upper class limit by **U.**

2) **Class-Intervals:** The difference between the lower limit and the upper limit of the class is known as the **class-interval**; **For example,** in the class 10 – 20, the class-interval is 10 (i.e., 20 –10). The formula to find the class-interval of a given problem is:

$$i = \frac{L - S}{k}$$

Where,

L = Largest item.

S = Smallest item .

k = The number of classes.

For example, if the marks of 50 students are varied between 10 and 80 and if we want to form 7 classes, then the class-intervals would be:

$$i = \frac{L - S}{k} : L = 80, S = 10, k = 7$$

$$= \frac{80 - 10}{7} = \frac{70}{7} = 10$$

Therefore, the class-intervals would be 10 – 20, 20 – 30, 30 – 40, 40 – 50, 50 – 60, 60 – 70, and 70 – 80.

Types of Class Intervals

There are three methods of classifying the data according to class intervals namely:

i) **Exclusive Method:** When the class intervals are so fixed that the upper limit of one class is the lower limit of the next class; it is known as the exclusive method of classification. **For example,** consider the following data, which are classified on the basis of this method:

Expenditure (₹)	No. of Families
0-5000	60
5000-10000	95
10000-15000	122
15000-20000	83
20000-25000	40
Total	400

From the above table, it is clear that the exclusive method ensures continuity of data as much as the upper limit of one class is the lower limit of the next class. In the example, there are 60 families whose expenditure is between ₹0 and ₹4999.99. A family whose expenditure is ₹5000 would be included in the class interval 5000-10000. This method is widely used in practice.

ii) **Inclusive Method:** In this method, the overlapping of the class intervals is avoided. Both the lower and upper limits are included in the class interval. This type of classification may be used for a grouped frequency distribution for discrete variable like members in a family, number of workers in a factory etc., where the variable may take only integral values. It cannot be used with fractional values like age, height, weight etc. **For example,** consider the following table which described the inclusive method

Class Interval	Frequency
5-9	7
10-14	12
15-19	15
20-29	21
30-34	10
35-39	5
Total	70

Thus, to decide whether to use the inclusive method or the exclusive method, it is important to determine whether the variable under observation in a continuous or discrete one. In case of continuous variables, the exclusive method must be used. The inclusive method should be used in case of discrete variable.

Scanned with CamScanner

iii) **Open End Classes:** A class limit is missing either at the lower end of the first class interval or at the upper end of the last class interval or both are not specified. The necessity of open end classes arises in a number of practical situations, particularly relating to economic and medical data when there are few very high values or few very low values which are far apart from the majority of observations.

The example for the open-end classes as follows:

Salary Range	No. of Workers
Below 2000	7
2000 - 4000	5
4000 - 6000	6
6000 - 8000	4
8000 and above	3

3) **Range:** The difference between largest and smallest value of the observation is called the Range and is denoted by 'R', i.e.,

R = Largest value – Smallest value

R = L – S

4) **Mid-Value or Mid-Point:** The central point of a class interval is called the mid value or mid-point. It is calculated by adding the upper and lower limits of a class and dividing the sum by 2.

$$\text{i.e., Mid-value} = \frac{L+U}{2}$$

For example, if the class interval is 20-30 then the mid-value is

$$\frac{20+30}{2} = 25$$

5) **Frequency:** Number of observations falling within a particular class interval is called frequency of that class. **For example,** Let us consider the frequency distribution of weights of persons working in a company.

Weight (in kgs)	Number of Persons
30-40	25
40-50	53
50-60	77
60-70	95
70-80	80
80-90	60
90-100	30
Total	420

In this example, the class frequencies are 25, 53, 77, 95, 80, 60, and 30. The total frequency is equal to 420. The total frequency indicates the total number of observations considered in a frequency distribution.

6) **Number of Class Intervals:** The number of class interval in a frequency is matter of importance. The number of class interval should not be too many. For an ideal frequency distribution, the number of class intervals can vary from 5 to 15. To decide the number of class intervals for the frequency distributive in the whole data, we choose the lowest and the highest of the values.

The difference between them will enable us to decide the class intervals. Thus the number of class intervals can be fixed arbitrarily keeping in view the nature of

problem under study or it can be decided with the help of Sturges' Rule. According to him, the number of classes can be determined by the formula

$$K = 1 + 3.322 \log_{10} N$$

Where N = Total number of observations.

log = Logarithm of the number.

K = Number of class intervals.

Thus, if the number of observation is 10, then the number of class intervals is

$K = 1 + 3.322 \log 10 = 4.322 \cong 4$

If 100 observations are being studied, the number of class interval is

$K = 1 + 3.322 \log 100 = 7.644 \cong 8$ and so on.

7) **Size of the Class Interval:** Since the size of the class interval is inversely proportional to the number of class interval in a given distribution. The approximate value of the size (or width or magnitude) of the class interval 'C' is obtained by using sturges rule as:

$$\text{Size of class interval(C)} = \frac{\text{Range}}{\text{Number of Class Interval}}$$

$$= \frac{\text{Range}}{1+3.322 \log_{10} N}$$

Where Range = (Largest Value – smallest value) in the distribution.

1.2.3.3. Cumulative Frequency Distribution

A cumulative distribution of frequencies shows the number of data items with values less than or equal to the upper class limit of each class. While a cumulative relative frequency distribution gives the proportion of the data items and a cumulative percentage frequency distribution shows the percentage of data items with values less than or equal to the upper class limit of each class.

Table 1.1: GCS Scores of Table Sharing the Cumulative Frequency

GCS Score	Frequency (Number of Patients)	Cumulative Frequency (Cumulative Number of Patients)
3	10	10
4	5	15
5	6	21
6	2	23
7	12	35
8	15	50
9	18	68
10	14	82
11	15	97
12	21	118
13	13	131
14	17	148
15	6	154

The cumulative frequency for each category tells us how many subjects there are in that category, and in all the lesser-valued categories in the table. **For example,** 35 patients had a GCS score of 7 or less.

Scanned with CamScanner

1.3. EXERCISE

1.3.1. Very Short Answer Type Questions

1) Define the biostatistics.
2) Write the uses of biostatistics.
3) What is frequency?
4) Write the formula to calculate the number of class intervals.

1.3.2. Short Answer Type Questions

1) What are the different types of statistics?
2) Explain the characteristics of statistics.
3) What do you mean by frequency distribution?
4) Why Statistics is considered as Statistical Methods

1.3.3. Long Answer Type Questions

1) Explain the scope/application of statistics.
2) Explain various type of frequency distribution with the suitable example.
3) Convert the following distribution into 'more than' frequency distribution:

Weekly Wages Less than	No. of Workers
20	41
40	92
60	156
80	194
100	201

4) Construct a frequency distribution table for the data on weights (in kg.) of 20 students of a class using intervals 30 – 35, 35 – 40 and so on.
40, 38, 33, 48, 60, 53, 31, 46, 34, 36, 49, 41, 55, 49, 65, 42, 44, 47, 48, 39.

5) Construct a continuous frequencies distribution (Take first class as 10-20) for the marks obtained by 50 students:

42	53	65	63	61	47	58	60	64	45	55	57	82	42	39	51	65	
55	33	70	50	52	53	45	45	25	36	59	63	39	65	30	45	35	
39	15	54	48	64	26	75	20	42	40	41	55	42	46	35	18		

6) From the following observations prepare a frequency distribution table in ascending order starting with 100-110 (exclusive method):

125	108	112	126	110	132	136	130	149	155
120	130	136	138	125	111	112	125	140	148
147	137	145	150	142	135	137	132	165	154

CHAPTER 2

Measures of Central Tendency

2.1. MEASURES OF CENTRAL TENDENCY

2.1.1. Meaning and Definition

Measure of central tendency is also known as **measure of central value** or **measure of location** or **average of first order.** It is a statistical measure and calculates the location or position of central point to explain the central tendency of the whole quantity of data.

According to Simpson and Kafka, "A measure of location is a typical value around which other figures congregate".

According to R.F. Fesher, "The inherent inability of the human mind to grasp in its entirely a large body of numerical data, compels us to seek relatively few constants that will describe the data".

According to Bowley. "Statistical constants enable us to comprehend in a single effort the significance of the whole".

According to M.R. Speigal, "Average is a value which is typical or representative of a set of data".

According to Clark and Sekkade, "Average is an attempt to find one single figure to describe whole of figure".

2.1.2. Importance of Central Tendency

1) **To Find Representative Value:** Measures of central tendency or averages give us one value for the distribution and this value represents the entire distribution. In this way averages convert a group of figures into one value.
2) **To Condense Data:** Collected and classified figures are vast. To condense these figures we use average. Average converts the whole set of figures into just one figure and thus helps in condensation.
3) **To Make Comparisons:** To make comparisons of two or more than two distributions, we have to find the representative values of these distributions. These representative values are found with the help of measures of the central tendency.
4) **Helpful in Further Statistical Analysis:** Many techniques of statistical analysis like Measures of Dispersion, Measures of Skewness, Measures of Correlation, and Index Numbers are based on measures of central tendency. That is why; measures of central tendency are also called as measures of the first order.

2.1.3. Properties of Good Measures of Central Tendency

1) It should be rigidly defined.
2) Its definition should be in the form of a mathematical formula.
3) It should be easy to understand and calculate.
4) It should remain unaffected by the extreme values.
5) It should be capable of further algebraic treatment.

Scanned with CamScanner

6) It should be based on all items in the series.
7) It should be capable of being used in further statistical computation or processing.
8) It should possess sampling stability.

2.1.4. Types of Average

In statistics, there are various types of measures of central tendency. Some of which can be broadly classified as follows:

1) Mathematical average, and 2) Positional average.

2.1.4.1. Mathematical Average

Mathematical averages are those averages which are calculated by taking into account the values of all the items in a series. When values of all the items in a series are considered in the calculation of average then it is called mathematical averages.

1) **Mean:** Mean is also known as Arithmetic Mean (A.M.). To calculate the mean, summate all the observations and divide it by the total number of observations. Mean is denoted by $\overline{X}$.

So, $$\overline{X} = \frac{\text{Sum of all the observations}}{\text{Number of observations}} = \frac{\sum X}{N}$$

2) **Geometric Mean:** The geometric mean of 'N' numbers is defined as the "N^{th} root of the product of 'N' numbers." If all the values of a series are multiplied and then N^{th} root of the product is extracted, then the geometric mean is obtained. Symbolically:

$$G.M. = \sqrt[N]{(X_1)\times(X_2)\times(X_3).....\times(X_n)}$$

3) **Harmonic Mean:** The harmonic mean is the reciprocals of averaged numbers. It is defined as the reciprocal of the arithmetic mean of the reciprocal of the individual observations. Under certain conditions, harmonic mean is a better measure of central tendency e.g., computation of average speed, average price, etc.

Mathematically, $$H.M. = \frac{N}{\frac{1}{X_1}+\frac{1}{X_2}+\frac{1}{X_3}+.....+\frac{1}{X_N}}$$

Where, N = Number of observations; X_N = Value of observation

2.1.4.2. Positional Average

Positional average is the average which depends on the position of the items, rather than the values of the items.

1) **Median:** When N observations are grouped and arranged in the sorting order (ascending or descending order) according to their values, then the central value of the observation is known as median. It is denoted by M or Me.
2) **Mode:** Mode is the value of the variable for which the frequency is maximum and it is denoted by Z or Mo.
3) **Quartiles:** The measure of central tendency which divides a group of data into four subgroups or parts, then it is called quartile. The three quartiles are denoted as Q_1, Q_2, and Q_3.
4) **Deciles:** If the values of the observation in a data set are arranged in ascending or descending order and divided it into ten equal parts by using nine points on the scale of observations, then it is called decile.
5) **Percentiles:** If the values of the observation in a data set are arranged in ascending or descending order and divided it into hundred equal parts by using ninety nine points on the scale of observations, then it is called percentile.

2.2. MEAN

2.2.1. Introduction

Mean is also known as **arithmetic mean (A.M.)**. To calculate the mean, summate all the observations and divide it by the total number of observations. Mean is denoted by $\overline{X}$.

So, $$\overline{X} = \frac{\text{Sum of all observations}}{\text{Number of observations}} = \frac{\sum X}{N}$$

According to W. I. King, "The arithmetic average may be defined as the sum or aggregate of a series of items divided by their number".

For example, the arithmetic mean of the four numbers 32, 40, 44 and 36 is given by,

$$\overline{X} = \frac{32+40+44+36}{4} = \frac{152}{4} = 38$$

2.2.2. Advantages of Mean

1) Simple to calculate and understand.
2) Some value is always determined, i.e., it is never indefinite.
3) Can be used in other algebraic calculations.
4) No need of sorting or arrangement (ascending or descending order).
5) It is stable and not affected by the variation of sampling.

2.2.3. Disadvantages of Mean

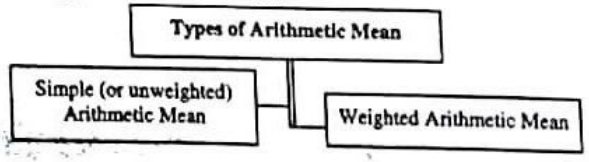
1) Mean is greatly affected by the extreme values. **For example,** mean of 3, 7 and 200 is 70. Here, no value is present near this 70; hence this average is of no use.
2) Sometimes mean may provide confusing impressions. **For example,** in a hospital, per day average number of patients admitted is 5.7. Here given information is useful but doesn't provide the actual item because some values are of no use when expressed in fraction or decimal.
3) The 'mean' cannot be predicted by just inspecting the sample item.
4) If a single value is missing, mean cannot be calculated.
5) In case of open-end classes, mean cannot be calculated.
6) Graphical representation of mean is not possible.

2.2.4. Application of Arithmetic mean

1) Arithmetic mean is used to measure the standard deviation.
2) Arithmetic mean is used in the construction of index numbers.
3) It is also used in the hypothesis testing.

2.2.5. Types of Arithmetic Mean

There are two types of arithmetic mean:



Scanned with CamScanner

2.2.6. Method of Calculation of Mean

The methods for calculating arithmetic mean vary according to various types of series such as:

1) Calculation of Arithmetic Mean - Individual Series
2) Calculation of Arithmetic Mean - Discrete Series (Ungrouped Data)
3) Calculation of Arithmetic Mean - Continuous Series (Grouped Data)

2.2.6.1. Calculation of Arithmetic Mean - Individual Series

The methods used are as follows:

1) **Direct Method:** The direct method of calculating arithmetic mean in the case of individual series involves the following steps:
   i) Add the various given values of the variable X and find the total value which is denoted by ΣX.
   ii) Divide this total value (ΣX) by the number of observations (N) i.e.,

$$\bar{X} = \frac{X_1 + X_2 + X_3 + \ldots + X_n}{N} = \frac{\Sigma X}{N}$$

Example 1: The weekly wage of 5 workers is as given below:
1350, 1400, 1450, 1370 and 1480

Find arithmetic mean.

Solution:

Computation of Arithmetic Mean

Serial Number	Weekly Wages (in Rupees)
1	1350
2	1400
3	1450
4	1370
5	1480
N = 5	ΣX = 7050

$$\bar{X} = \frac{\sum X}{N} = \frac{7050}{5} = ₹1410$$

2) **Short-Cut Method:** With the use of short cut method also, arithmetic mean can be calculated. The calculation of arithmetic mean by short cut method includes the following steps:
   i) Any value may be taken as an assumed mean of the data. This assumed value is also known as working mean or arbitrary average (A = Assumed mean).
   ii) Subtract assumed mean from each value of the observation ($d = X - A$).
   iii) All the deviations are added and it is denoted by (Σd).
   iv) Apply the formula:

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

Where,
$\bar{X}$ = Arithmetic mean; A = Assumed mean
$\sum d$ = Sum of the deviations; N = Number of items

Example 2: Calculate mean from the following data

Years	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010
Price of Rice (in ₹)	40	50	55	60	58	60	75	35	45	52

Solution: Let the assumed mean be A = 50

Years	Price of Rice (X)	d = (X - 50)
2001	40	-10
2002	50	0
2003	55	5
2004	60	10
2005	58	8
2006	60	10
2007	75	25
2008	35	-15
2009	45	-5
2010	52	2
N = 10		Σd = 50

$$\bar{X} = A + \frac{\sum d}{N} = 50 + \frac{50}{10} = 50 + 5 = ₹55$$

2.2.6.2. Calculation of Arithmetic Mean-Discrete Series (Ungrouped Data)

The methods used are as follows:

1) **Direct Method:** In the discrete series, the sum of items is determined by multiplying each value with the respective frequency. The values got after multiplication are totalled up. To find the arithmetic mean, total value is divided by the total number of items. Following are the steps which are involved in the calculation of mean:
   i) Multiply each item by its frequency, denoted by (fX).
   ii) Add all the fX, denoted by (ΣfX).
   iii) ΣfX is divided by the total number of items.

The formula is $\bar{X} = \frac{\sum fX}{\sum f}$

Where, $\bar{X}$ = Arithmetic mean;
ΣfX = The sum of products;
Σf = Total number of items.

Example 3: Find average wages of 10 workers:

Daily Wage (in ₹)	4	6	10	11	14	Total
No. of Workers	2	1	4	2	1	10

Solution:

Daily Wage (X)	No. of Workers (f)	fX
4	2	8
6	1	6
10	4	40
11	2	22
14	1	14
Total	Σf = 10	ΣfX = 90

$\therefore$ Arithmetic Mean (Average Wage) = $\frac{\sum fX}{\sum f} = \frac{90}{10} = ₹9.00$

2) **Short-Cut Method:** For calculating the arithmetic mean by short-cut method in a discrete series the following steps are taken:
   i) Any value may be taken as an assumed mean of the data. This assumed value is also known as working mean or arbitrary average (A = Assumed mean). For

Scanned with CamScanner

making the calculation easy, the assumed mean should be close to the middle of the frequency distribution.

ii) Subtract assumed mean from each value of the observation ($d_x = X - A$).

iii) Each value of deviation is multiplied by its respective frequency, denoted by (fd_x).

iv) Following is the formula which is used for the calculation of arithmetic mean:

$$\bar{X} = A + \frac{\sum fd_x}{\sum f} \quad \text{or} \quad A + \frac{\sum fd_x}{\sum N}$$

Where, $\bar{X}$ = Arithmetic Mean

A = Assumed Mean

Σfd_x = Total of deviations multiplied with the respective frequencies

Σf = Total of frequencies (N)

Example 4: From the following frequency distributions find out the mean weight of the 100 persons:

Weight (in kg.)	64	65	66	67	68	69	70	71	72	73
No. of Persons	15	13	18	5	20	11	7	6	3	2

Solution: Let assumed mean (A) = 68;

Short-cut Method

Weight in Kg. (X)	No. of Persons (f)	Deviation (d_x) = (X – A)	fd_x
64	15	– 4	– 60
65	13	– 3	– 39
66	18	– 2	– 36
67	5	– 1	– 5
68 (A)	20	0	0
69	11	1	11
70	7	2	14
71	6	3	18
72	3	4	12
73	2	5	10
	$\Sigma f = 100$		$\Sigma fd_x = -75$

$$\bar{X} = A + \frac{\sum fd_x}{\sum f} = 68 + \frac{(-75)}{100} = 68 - 0.75 = 67.25 \text{ Kg.}$$

2.2.6.3. Calculation of Arithmetic Mean - Continuous Series (Grouped Data)

The methods used are as follows:

1) **Direct Method:** For calculating the arithmetic mean in a continuous series the following steps are taken:

i) First mid value of each class-interval is determined by adding the lower and upper limit of each class and dividing the total by two. **For example,** in a class interval say 0 – 10, the mid value is 5.

$$\left(\frac{0+10}{2} = 10/2 = 5\right)$$ (This is denoted by X).

ii) Multiply these mid values of each class with the respective frequency of each class. In other words X will be multiplied by f.

iii) Add up all the products and obtain ΣfX.

iv) ΣfX is divided by the sum of the frequencies i.e., Σf.

v) Apply the formula: $\bar{X} = \frac{\sum fX}{\sum f}$

Example 5: From the following find out the mean profits

Profits per Share (₹)	100-200	200-300	300-400	400-500	500-600	600-700	700-800
Number of Shares	12	20	18	30	32	26	22

Solution:

Calculation of Mean

Profits	Mid-Point (X)	No. of Shares (f)	fX
100-200	150	12	1800
200-300	250	20	5000
300-400	350	18	6300
400-500	450	30	13500
500-600	550	32	17600
600-700	650	26	16900
700-800	750	22	16500
		$\Sigma f = 160$	$\Sigma fX = 77600$

$$\bar{X} = \frac{\sum fX}{\sum f} = \frac{77600}{160} = 485$$

The average profit is ₹485.

2) **Step-Deviation Method:** For calculating the arithmetic mean by step-deviation method in case of grouped series the following steps are taken:

i) Find the mid-point of each group or class, denoted by X.

ii) Any value may be taken as an assumed mean of the data (A = Assumed mean).

iii) Subtract assumed mean from the mid-point of each class, denoted by $d_x = (X - A)$.

iv) Deviations (d_x) are divided by their common factor (i) and step deviation is determined for each class-interval.

$$d' = \frac{d_x}{i}.$$

Thus, $d' = \frac{X - A}{i}$

v) The step deviation of each class is multiplied by respective frequency of that class to find the total value $\Sigma fd'$.

vi) Find the total frequency, $N = \Sigma f$.

Use the formula:

$$\bar{X} = A + \frac{\sum fd'}{\sum f} \times i$$

Where,

$\bar{X}$ = Arithmetic mean,

A = Assumed mean,

$\Sigma fd'$ = Total of product of step-deviations and frequencies; Σf = Total number of frequencies;

i = Common factor in x or in d_x

Scanned with CamScanner

Example 6: Calculate the mean from the following table

Weights of Children (in kg)	0–10	10–20	20–30	30–40	40–50	50–60	60–70	70–80	80–90
Number of Children	4	12	8	21	32	28	10	3	2

Solution: Let assumed mean (A) = 55, let i = 10 (Common Factor)

C.I (Weight in kg.)	Mid Value (X)	Frequency (f)	$d_x = X - A$	$d' = d_x/10$	fd'
0–10	5	4	– 50	– 5	– 20
10–20	15	12	– 40	– 4	– 48
20–30	25	8	– 30	– 3	– 24
30–40	35	21	– 20	– 2	– 42
40–50	45	32	– 10	– 1	– 32
50–60	55	28	0	0	0
60–70	65	10	10	1	10
70–80	75	3	20	2	6
80–90	85	2	30	3	6
Total		$\Sigma f = 120$			$\Sigma fd' = -144$

$$A.M. = A + \frac{\sum fd'}{\sum f} \times i = 55 + \frac{(-144)}{120} \times 10$$

$= 55 - 12$ or $= 43$ Kg.

2.2.7. Weighted Arithmetic Mean

In the economic studies, weighted arithmetic mean plays a very important role. In the calculation of weighted arithmetic mean, items are assigned weights according to their relative importance in the group. The relative importance of various items in a distribution may be different, some items may be relatively more important than others. **For example,** to calculate mean expenditure of a family, it would be wrong to give equal importance to different items of family expenditure. The family may give different importance to the various items depending upon how much monthly expenditure is incurred on those items. They may give more importance to food and less importance to entertainment or clothes. Hence, the arithmetic mean should be calculated according to their relative importance.

Methods of Calculation of Weighted Arithmetic Mean

1) **Direct Method:** The following steps are taken in the calculation of weighted arithmetic mean by direct method:
   i) Multiply size of the variable (X) and their respective weights and add up the products.
   ii) Divide the total by sum of the weights.
   iii) Following formula is used for the calculation of weighted arithmetic mean:

$$\bar{X}_w = \frac{\sum WX}{\sum W}$$

Where, $\bar{X}_w$ = Weighted arithmetic mean

W = Weights,

X = The variable

2) **Short-Cut Method:** The following steps are taken in the calculation of weighted arithmetic mean by short-cut method:
   i) Any value may be taken as an assumed weighted mean of the data (A_w = Assumed mean)

   ii) Subtract assumed mean from each value of the observation ($d_x = X - A_w$)
   iii) Each value of deviation is multiplied with its respective weight, denoted by Wd_x and then added up, $\sum (Wd_x)$
   iv) Following is the formula which is used in the calculation of weighted arithmetic mean:

$$\bar{X}_w = A_w + \frac{\sum Wd_x}{\sum W}$$

Where,

$\bar{X}_w$ = Weighted arithmetic mean;

A_w = Assumed weighted mean

ΣWd_x = Total of deviations multiplied with the respective weights;

ΣW = Total of weights

Example 7: An examination was held to decide the award of a scholarship. The weights given to the various subjects were different. Only three applicants for the scholarship obtained over 50% marks in each subject. The marks were as follows:

Subject	Weight	Marks of A	Marks of B	Marks of C
Statistics	4	65	58	66
Accountancy	3	64	65	72
Economics	2	58	57	58
Mercantile Law	1	72	79	54

Of the candidates the one getting the highest marks is to be awarded the scholarship, who should get it?

Solution:

Subject	Weight W	Marks of A		Marks of B		Marks of C	
		X_1	WX_1	X_2	WX_2	X_3	WX_3
Statistics	4	65	260	58	232	66	264
Accountancy	3	64	192	65	195	72	216
Economics	2	58	116	57	114	58	116
Mercantile Law	1	72	72	79	79	54	54
Total	$\Sigma W = 10$		$\Sigma WX_1 = 640$		$\Sigma WX_2 = 620$		$\Sigma WX_3 = 650$

Weighted Mean of A, $\bar{X}_{w_A} = \frac{\sum WX_1}{\sum W} = \frac{640}{10} = 64$ Marks

Weighted Mean of B, $\bar{X}_{w_B} = \frac{\sum WX_2}{\sum W} = \frac{620}{10} = 62$ Marks

Weighted Mean of C, $\bar{X}_{w_C} = \frac{\sum WX_3}{\sum W} = \frac{650}{10} = 65$ Marks

The Weighted Mean of 'C' is the highest; hence he is entitled for scholarship.

2.2.8. Combined Mean/Grouped Mean

If the means of two or more than two sub-groups or different samples from the same universe along with their numbers are given, then a combined mean can be calculated with the help of following formula:

$$\text{Combined Mean } (\bar{X}) = \frac{N_1\bar{X}_1 + N_2\bar{X}_2 + N_3\bar{X}_3 \ldots\ldots + N_n\bar{X}_n}{N_1 + N_2 + N_3 \ldots\ldots + N_n}$$

Scanned with CamScanner

N_1, N_2, N_3,..... are the frequencies of different groups and $\bar{X}_1, \bar{X}_2, \bar{X}_3$..... are the arithmetic means of these groups.

Example 8: An analysis of the average speed of the buses in two transport companies A and B, gives the following results:

	Transport Company A	Transport Company B
Number of Buses	595	645
Average Speed (Km/h)	52.5	47.5

Find the combined average speed.

Solution: Combined Average Speed in the two transport companies would be,

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

Where, N_1 and N_2 = the number of items in the two series

$\bar{X}_1$ and $\bar{X}_2$ = the mean of the two series respectively

$\bar{X}_{12}$ = the combined mean of the two series.

Given, $N_1 = 595$ and $N_2 = 645$; $\bar{X}_1 = 52.5$ and $\bar{X}_2 = 47.5$

Substituting the values,

$$\text{Combined Average Speed} = \frac{(595\times 52.5)+(645\times 47.5)}{595+645} = \frac{61875}{1240} = ₹49.9$$

2.3. MEDIAN

2.3.1. Introduction

When N observations are grouped and arranged in the sorting order (ascending or descending order) according to their values, then the central value of the observation is known as median. It is denoted by M or Me.

So,

$$M = \left(\frac{N+1}{2}\right)^{th} \text{ observation.}$$

According to Prof. L R. Conner, "The median is that value of the variable which divides the group into two equal parts, one part comprising of all values greater and the other all values less than the median".

According to Croxton and Cowden, "The median is that value which divides a series so that one half or more of the items are equal to or less than it and one half or more of the items are equal to or greater than it."

2.3.2. Advantages of Median

1) It is unaffected by the extreme values.
2) In case of individual observation and discrete series, it is easy to understand and calculate.
3) It can be used in other algebraic calculations and also useful during calculation of mean deviation.

4) After sorting the values of variable, it can be located by inspection.
5) It can also be measured by using 'graphical representation'.
6) Median of open-end classes can be measured.
7) Median is clearly definite in nature i.e., it is always clearly defined.

2.3.3. Disadvantages of Median

1) Sorting (ascending and descending order) is necessary for the calculation.
2) It cannot be used to measure the combined mean of two or more groups.
3) When sampling is fluctuated then median also gets fluctuated (affected more than mean).
4) It is not based on all the data samples. It is a positional average.

2.3.4. Applications of Median

The median can be used to understand the features of a data set when:

1) Observations are qualitative in nature.
2) Extreme values are present in the data set.
3) A fast estimate of an average is desired.

2.3.5. Method of Calculation of Median

1) Calculation of Median – Individual Series
2) Calculation of Median – Discrete Series (Ungrouped Data)
3) Calculation of Median – Continuous Series (Grouped Data)

2.3.5.1. Calculation of Median – Individual Series

In case of individual series, the following steps are taken for calculating the median:

1) If the given observations are not arranged in ascending or descending order of magnitude then it has to be arranged for calculation.
2) Locate the middle value:
   i) There will be a single value in the middle which is selected as median when the number of observation N is odd. Mathematically

$$\text{Median} = \text{Value of } \left(\frac{N+1}{2}\right)^{th} \text{ items}$$

Example 9: Find out the median of the following items:

X : 10, 15, 9, 25, 18

Solution:

Computation of Median

S. No.	Size of Item Ascending Order (X)		Size of Item Descending Order (X)
1)	9		25
2)	10		18
3)	15	← [Median] →	15
4)	18		10
5)	25		9

$$\text{Median} = \text{Size of } \left(\frac{N+1}{2}\right)^{th} \text{ item} = \text{Size of } \left(\frac{5+1}{2}\right)^{th} \text{ item} = 3^{rd} \text{ item} = 15$$

In the above example the number of items was odd so there was no difficulty in locating the middle item and its value.

Scanned with CamScanner

ii) If N is even, there will be two middle values, $\left(\frac{N}{2}\right)^{th}$ and $\left(\frac{N}{2}+1\right)^{th}$. The arithmetic mean of these two middle values is taken as median. That is

$$\text{Median} = \frac{\left(\frac{N}{2}\right)^{th}\text{ value} + \left(\frac{N}{2}+1\right)^{th}\text{ value}}{2}.$$

Example 10: Find out the median of the following items:

X : 12, 18, 9, 25.

Solution:

Computation of Median

S. No.	Size of Item Ascending Order (X)
1)	9
2)	12
3)	18
4)	25

Here, N = 4 which is even, so

$$\text{Median} = \frac{\left(\frac{N}{2}\right)^{th}\text{ value} + \left(\frac{N}{2}+1\right)^{th}\text{ value}}{2} = \frac{\left(\frac{4}{2}\right)^{th}\text{ value} + \left(\frac{4}{2}+1\right)^{th}\text{ value}}{2}$$

$$= \frac{2^{nd}\text{ value} + 3^{rd}\text{ value}}{2}$$

$$= \frac{12+18}{2} = \frac{30}{2} = 15$$

2.3.5.2. Calculation of Median – Discrete Series (Ungrouped Data)

In case of discrete series, the following steps are taken for calculating the median:

1) If the given observations are not arranged in ascending or descending order of magnitude then it has to be arranged for calculation.
2) Calculate the total frequency, i.e., $\sum f = N$.
3) Calculate the cumulative frequency (in general, less than type, data arranged in ascending order).
4) Locate the median by using the following formula:

$$\text{Median}(M) = \left(\frac{N+1}{2}\right)^{th}\text{ value,}$$

(Where, $N = \sum f$)

The value, for which the cumulative frequency includes $\left(\frac{N+1}{2}\right)^{th}$ value, is selected as median.

Example 11: To find the median of the following distribution:

X	1	2	3	4	5	6
f	8	12	16	19	20	25

Solution:

X	f	Cumulative frequency
1	8	8
2	12	20
3	16	36
4 ← Median	19	55 ← Cumulative frequency of median
5	20	75
6	25	100
	N = 100	

Now median = Value of $\left(\frac{N+1}{2}\right)^{th}$ item = Value of $\left(\frac{100+1}{2}\right)^{th}$ item or, Value of 50.5^{th} item. This value lies in cumulative frequency (55) for the value 4. So, Median = 4.

2.3.5.3. Calculation of Median – Continuous Series (Grouped Data)

In case of continuous series, the following steps are taken for calculating the median:

1) If the given series is in inclusive form then it has to be converted in exclusive form.
2) Calculate the total frequency,

   $N = \sum f$ and median number $\left(\frac{N}{2}\right)^{th}$ value.
3) If the data is arranged in ascending order, then calculate 'less than' type cumulative frequencies.
4) Calculate median class with the use of $m = \frac{N}{2}$. The class in which median falls is called the median class.
5) Median can be calculated with the use of following formula:

$$\text{Median}(M) = L_1 + \frac{m-c}{f} \times i \quad \text{or} \quad M = L_1 + \frac{m-c}{f} \times (L_2 - L_1)$$

Where, M = median, m = median number = N/2,
L_1 = lower limit of the median class,
L_2 = upper limit of the median class,
f = frequency of the median class,
$i = L_2 - L_1$ = width of the median class (Class size).
c = cumulative frequency (less then type) of the class preceding to the median class,

Example 12: Find the median and median class of the following data:

Class-Interval	15-25	25-35	35-45	45-55	55-65	65-75
Frequency	5	10	18	15	0	3

Solution:

Class Interval	Frequency	Cumulative Frequency
15 – 25	5	5
25 – 35	10	15 ← 'c' cumulative frequency of the class preceding the median class
35 – 45 ← M. Class	18 ← f	34 ← Cumulative frequency of median class
45 – 55	15	49
55 – 65	0	49
65 – 75	3	52
	N = 51	

Scanned with CamScanner

Median number, $m = \frac{N}{2} = \frac{51}{2} = 26^{th}$ element, and this value lie in cumulative frequency (34) and class interval is (35 – 45). So the median class is (35– 45).

Here, m = 26, c = 15, f = 18 and $L_2 - L_1 = 10$.

Hence, $M = L_1 + \frac{m-c}{f} \times (L_2 - L_1) = 35 + \frac{26-15}{18} \times 10$

$= 35 + \frac{11}{18} \times 10 = 35 + \frac{110}{18} = 35 + 6.11 = 41.11$

Required median is 41.11 and median class is (35 – 45).

2.4. MODE

2.4.1. Introduction

Mode is the value of the variable for which the frequency is maximum and it is denoted by Z or Mo.

According to Zizek, "Mode is the value occurring most frequently in a series (or group) of items and around which other items are distributed most densely".

According to Croxton and Cowden, "The mode of a distribution is the value at the point around which the items tend to be most heavily concentrated. It may be regarded as the most typical of a series of values".

2.4.2. Advantages of Mode

1) Easy to understand and calculate.
2) Can be easily found out by using inspection method.
3) It is an actual value, which most frequently occurs in the series.
4) Not affected by extreme values.
5) It is simple and accurate and can be measured in an open end class-interval without determining the class limits.

2.4.3. Disadvantages of Mode

1) It is ill defined and in few cases it is not possible to find a definite value.
2) Not a good representative because it is not based on all observations.
3) In further algebraic calculation it is of no use.
4) It may be possible that there is no mode or more than one mode present in the sample.
5) Fluctuation in sampling effects is made to a higher degree than mean or median.
6) Before calculating the mode, data sorting (ascending or descending order) is necessary.

2.4.4. Methods of Calculation of Mode

1) Calculation of Mode – Individual Series
2) Calculation of Mode – Discrete Series (Ungrouped Data)
3) Calculation of Mode – Continuous Series (Grouped Data)

2.4.4.1. Calculation of Mode – Individual Series

If the data is not arranged then it has to be arranged in increasing order. Now find the value occurring the maximum number of times. This value is called the 'mode' of the given data.

In case, two or more values consist of the similar maximum frequency, one cannot say which the modal value is and then the mode is said to be **ill-defined**. Such a series is also called as **bi-modal or multi-modal.**

Example 13: Calculate mode from the following data of income earned by 10 employees:

S. No.	Income (in ₹)	S. No.	Income (in ₹)
1	10	6	27
2	27	7	20
3	24	8	18
4	18	9	15
5	27	10	32

Solution:

Calculation of Mode

Income (in ₹)	Number of Employees
10	1
15	1
18	2
20	1
24	1
27	3
32	1
Total	**10**

Since the item 27 occurs the maximum number of times, i.e. 3, hence the modal income is 27.

Note: The process of determining mode in case of individual observations essentially involves grouping of data.

2.4.4.2. Calculation of Mode – Discrete Series (Ungrouped Data)

In the discrete series the mode is the value which has highest frequency. If two or more values have the same frequency or distribution of frequency is irregular, then mode cannot be determined by the method of inspection.

In this situation, a grouping table and an analysis table is to be prepared to find out the value of mode. First we prepare grouping table and then an analysis table. The following steps are taken for calculating the mode:

1) A grouping table is prepared with 6 columns.
2) The size of the item is written in the margin.
3) Frequencies of the respective items are written in column 1 and highest frequency is marked.
4) In column 2, the frequencies are grouped in twos by adding 1st and 2nd, 3rd and 4th, 5th and 6th and so on and highest total is marked.
5) In column 3, leaving the first frequency in column 1and the rest frequencies are grouped in twos by adding 2nd and 3rd, 4th and 5th, 6th and 7th and so on and highest total is marked.

Scanned with CamScanner

6) In column 4, starting from the first frequency of column 1, the frequencies are grouped in threes by adding 1st, 2nd and 3rd; 4th, 5th and 6th; 7th, 8th, and 9th; and so on and highest total is marked.

7) In column 5, starting from the second frequency (leaving the first frequency) of column 1, frequencies are grouped in threes by adding 2nd, 3rd and 4th; 5th, 6th and 7th; 8th, 9th and 10th and so on and highest total is marked.

8) In column 6, starting from the second frequency (leaving the first two frequencies) of column 1, frequencies are grouped in threes by adding 3rd, 4th and 5th; 6th, 7th and 8th so on and highest total is marked.

9) After completing the grouping of frequencies table, an analysis table is prepared to find out the item (group) which appears as the highest frequency. The same steps are applied in the case of continuous series.

Example 14. Compute the mode from the following data:

Size	2	3	4	5	6	7	8	9	10	11	12	13
Frequency	3	8	10	12	16	14	10	8	17	5	4	1

Solution: Due to irregular distribution of frequencies, we use the method of grouping to decide which one may be considered as maximum frequency.

Grouping Table

Value of the Items (m)	Frequency or Sum of Frequencies					
	1	2	3	4	5	6
2	3	11	18	21	30	38
3	8					
4	10	22	28	42	40	32
5	12					
6	16	30	24	35	30	26
7	14					
8	10	18	25	10		
9	8					
10	17	22	9			
11	5					
12	4	5				
13	1					

Analysis table is as follows:

Analysis Table

Column No.	Size of Items Containing Maximum Frequency											
	2	3	4	5	6	7	8	9	10	11	12	13
1									1			
2					1	1						
3				1	1							
4				1	1	1						
5					1	1	1					
6			1	1	1							
No. of Items	0	0	1	3	5	3	1	0	0	0	0	0

Since 6 occurs maximum number of times, hence the value 6 is the mode.

2.4.4.3. Calculation of Mode – Continuous Series (Grouped Data)

In a continuous series, determination of mode requires one step more than that used for discrete series. Firstly, one has to find the modal class by observation, or in case of 'irregular distribution' the modal class is determined by preparing grouping table and analysis tables. Finally, mode is obtained by following formula:

$$\text{Mode or } Z = L_1 + \frac{f_m - f_1}{2f_m - f_1 - f_2} \times i$$

Where,

L_1 = lower limit of the modal class,

f_m = frequency of the modal class

f_1 = frequency of the class preceding the modal class,

f_2 = frequency of the class succeeding the modal class

$i = (L_2 - L_1)$ = Width of modal Class (class size)

Example 15: Following is the frequency distribution of yearly expenditure of 50 families:

Yearly Expenditure (in thousand)	50-55	55-60	60-65	65-70	70-75	75-80
No. of Families	3	8	14	20	16	2

Find the modal value.

Solution: Here the class-interval (65 – 70) has the maximum frequency, i.e., 20, therefore, modal class is (65 – 70).

Now we can calculate the modal value of using the formula, $\text{Mode} = L_1 + \frac{f_m - f_1}{2f_m - f_1 - f_2} \times i$

Here, $L_1 = 65$, $f_m = 20$, $f_1 = 14$, $f_2 = 16$ and $i = 5$;

$$\therefore \quad \text{Mode} = 65 + \frac{20-14}{2\times20-14-16} \times 5 = 65 + \frac{30}{10}$$

Or Modal value = 65 + 3 = ₹68 thousand

Note: Generally, modes are used for nominal scores, medians for ordinal scores, and means for interval scores.

2.5. RELATION BETWEEN MEAN, MEDIAN AND MODE

The empirical relation between Mean, Median and Mode is,

Mode = 3 Median – 2 Mean.

Karl Pearson has expressed this relationship as follows:

Mode = Mean – 3(Mean- Median)

Mode = 3 Median – 2 Mean

Or, **Median = Mode + 2/3(Mean - Mode)**

Example 16: Calculate the Median when Mean and Mode of Distribution are 38.6 and 32.6 respectively.

Solution: Given, Mean = 38.6 and Mode = 32.6.

$$\text{Median} = \text{Mode} + \frac{2}{3}[\text{Mean} - \text{Mode}] = 32.6 + \frac{2}{3}[38.6 - 32.6] = 32.6 + \frac{2}{3}[6] = 32.6 + 4$$

Median = 36.6

Scanned with CamScanner

Example 17: For a given set of values of a variable, the mean and mode are 28.16 and 24 respectively, calculate the median.

Solution: Mean = 28.16
Mode = 24

$$\text{Median} = \text{Mode} + \frac{2}{3}[\text{Mean} - \text{Mode}] = 24 + \frac{2}{3}[28.16 - 24] = 24 + \frac{2}{3}[4.16] = 24 + 2.77$$

Median = 26.77

2.6. COMPARISON OF MEAN, MEDIAN AND MODE

Table 2.1: shows the difference between Mean, Median and Mode:

Table 2.1: Difference between Mean, Median and Mode

Mean	Median	Mode
It is highly affected by extreme values.	It is not affected by extreme values as in case of mean.	It is also not affected by extreme values as in case of mean.
It is based on all the observations.	It is not based on all observations.	It is also not based on all observations.
It is not necessary to arrange the observations in ascending or descending order.	It is necessary to arrange the data in ascending or descending order of magnitude.	It is also necessary to arrange the data in ascending or descending order of magnitude.
It is capable of further algebraic treatments.	It is also capable of further algebraic treatment.	It is not capable of further algebraic treatment.
It is rigidly defined and determined.	It is also rigidly defined and determined.	It is ill defined.
It can hardly be located by inspection.	It can be located by inspection, after arranging the data in order of magnitude.	It can also be found out by inspection.

2.7. PHARMACEUTICAL EXAMPLES

Example 18: The blood pressure (mmHg) reduction in 6 patients are measured by administration after 4 hours. The standard dose of a novel antihypertensive agent is illustrated in **table 2.2.**

Table 2.2: Effect of an Antihypertensive Drug on Blood Pressure Lowering in Six Patients

Patient Number	Reduction in Blood Pressure (mmHg)
1	20
2	25
3	21
4	34
5	41
6	37

Determine the mean reduction or drop in blood pressure reduction in the 6 patients.

Solution: Putting the values from **table 2.2** into the arithmetic equation, we obtain:

$$\text{Mean} = \frac{\sum_{j=1}^{N} X_j}{N} = \frac{(20+25+21+34+41+37)}{6} = \frac{178}{6} = 29.67 \text{ mmHg}$$

Example 19: The effect of a defined dose of a painkiller to suppress pain following a painful stimulus was calculated in 20 volunteers using a visual analogue scale. The results are illustrated in **table 2.3.**

Table 2.3: Recorded Assessment of Pain by 20 Volunteers

No. of Volunteers	Pain Assessment by Volunteers
2	3 (Extreme Pain)
12	2 (Moderate Pain)
6	1 (Slight Pain)

Determine the mean of the pain assessment by the 20 volunteers.

Solution: This example defines that the three sub-groups describe different clinical effects and are not equal in magnitude (weighting). The following equation can be used to calculate the weighted average:

$$\text{Mean} = \frac{\sum_{j=1}^{N} w_j X_j}{N}$$

Where, X_j = Data from j = 1 to j = N, N = Number of data
w_j = Weighting (frequency) of each group or series

$$\text{Mean} = \frac{\sum_{j=1}^{N} w_j X_j}{N} = \frac{(2\times3)+(12\times2)+(6\times1)}{20} = \frac{36}{20} = 1.8$$

Example 20: The adherence of the opportunistic pathogen Candida albicans to 10 Buccal Epithelial Cells (BEC) in vitro was calculated and is illustrated in **table 2.4.**

Table 2.4: Number of Blastospores of Candida Albicans Adherent to a Defined Number of Buccal Epithelial Cells, as Examined using an *in vitro* Method

Epithelial Cell Number	Number of Adherent C. Albicans
1	2
2	0
3	0
4	6
5	4
6	24
7	9
8	6
9	1
10	0

Find out the mean and median values for the adherence profile of this pathogen.

Scanned with CamScanner

Solution: Calculation of the Mean

$$\text{Mean} = \frac{\sum_{i=1}^{n} X_i}{N} = \frac{(2+0+0+6+4+24+9+6+1+0)}{10} = 5.2$$ Blastospores of C. albicans per buccal epithelial cell.

Calculation of Medium

First we arrange the data in ascending order as follows:
0, 0, 0, 1, 2, 4, 6, 6, 9, 24

Find the mean of values (central value) of position 5 and 6. The calculated median is (2 + 4)/2= 3 adherent blastospores of C. albicans per buccal epithelial cell.

Example 21: Using chromatographic methods, it is determined the concentration of therapeutic (mg/mL) agent in 10 vials of parenteral product.

Table 2.5: Concentration of a Therapeutic Agent in 10 Vials of a Product

Vial Number	Concentration of therapeutic agent(mg/mL)
1	200
2	205
3	205
4	201
5	199
6	195
7	202
8	205
9	205
10	207

Find out the mode of the observed concentrations.

Solution: 205 mg/mL is the most common observed value in the above data set; hence this is the value of mode.

Example 22: Determine the mean weights of the following tablets:
348mg, 350mg, 341mg, 356mg, 350mg, 359mg, 344mg, 352mg.

Solution: Calculation of Mean

$$\text{Mean} = \frac{348+350+341+356+350+359+344+352}{8} = 350$$

Thus, Mean = 350mg.

Example 23: The respiratory rates of 10 infants suspected to be suffering from acute respiratory infection are given as follows:
53, 52, 55, 48, 54, 59, 46, 58, 49, 60

Calculate the measures of central tendency.

Solution:

1) $\text{Mean} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{53+52+55+48+54+59+46+58+49+60}{10} = \frac{534}{10} = 53.4$

2) Median $= \left(\frac{n+1}{2}\right)^{th}$ in arranged series

Arranging the data in an order:
46, 48, 49, 52, 53, 54, 55, 58, 59, 60

$$\text{Median} = \left(\frac{10+1}{2}\right)^{th} \text{value}$$

$$= \left(\frac{11}{2}\right)^{th} \text{value} = 5.5^{th}$$

i.e., Average of 5th and 6th value:

$$= \frac{53+54}{2} = 53.5$$

3) Mode = Mode does not exist.

Mode is the values which is mostly repeated or have the highest frequency. In this given problem, there is not repetition of any value, hence no mode exist.

Example 24: The total weight (kg) gain of 15 women at the end of 9th month of pregnancy are as follows:
06, 08, 07, 10, 06, 07, 07, 08, 09, 11, 10, 10, 12, 14, 12

Determine the mean, median and also define the limitations of mean.

Solution:

1) $\text{Mean} = \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{137}{15} = 9.1333$

2) In this series, the most middle value is the median. First we arrange the series in ascending order as follows:
06, 06, 07, 07, 07, 08, 08, 09, 10, 10, 10, 11, 12, 12, 14

$$\text{Median} = \left(\frac{n+1}{2}\right)^{th} \text{value} = \left(\frac{15+1}{2}\right)^{th} \text{value} = \left(\frac{16}{2}\right)^{th} \text{value} = 8^{th} \text{value}$$

Median = 9

3) Mode = Most repeated value

The most repeat values are 7(three times) and 10(three times). Therefore, this is a "Bimodel" distribution. Both 7 and 10 are the modes.

2.8. EXERCISE

2.8.1. Very Short Answer Type Questions

1) What do you mean by measures of central tendency?
2) What is mean?
3) What is mode?
4) Write the application of arithmetic mean.
5) Define median with suitable example.

Scanned with CamScanner

2.8.2. Short Answer Type Questions

1) Write the properties of good measures of central tendency.
2) Explain the various type of average.
3) Write the advantage and disadvantages of mean.
4) Explain the relationship between mean, median and mode with example.
5) Write the advantages and disadvantages for mode.

2.8.3. Long Answer Type Questions

1) Explain the types of arithmetic mean with suitable example.
2) Compare mean, median and mode.
3) Discuss about the calculation of mode and also give an example.
4) What do you mean by method of calculation of median? And write the example.
5) Find the arithmetic mean of following series:

X	10	20	30	40	50
f	7	9	5	6	3

[Ans: Mean Salary: 4666.7]

6) Find out the median of the following items:
   X : 12, 18, 9, 25.
   [Ans: median =15]

7) The prices (in ₹) of the shares of a company over last 10 weeks were as follows:
   232 235 229 232 240 260 251 243 238 218

   Find average price of share using:
   i) Mean
   ii) Median
   iii) Mode

   [Ans: Mean = 237.8, Median = 236.5, Mode = 232]

8) The monthly income of 5 persons is 132, 140, 144, 136 and 148. Find arithmetic mean.

   [Ans: Mean = 140]

9) Find the median and median class of the following data:

Class-Interval	15-25	25-35	35-45	45-55	55-65	65-75
Frequency	5	10	18	15	0	3

   [Ans: Median = 41.11]

10) For the following frequency table calculate mean, median and mode:

Weekly Rent (in ₹)	No. of Persons Paying the Rent	Weekly Rent (in ₹)	No. of Persons Paying the Rent
200-400	6	1,200-1,400	15
400-600	9	1,400 -1,600	10
600-800	11	1,600 -1,800	8
800-1,000	14	1,800-2,000	7
1,000-1,200	20		

    [Ans: Mean = Med = 1,100; Mode = 1,109.09]

11) The yearly expenditure of 6 families on vegetables is 14000, 15000, 14200, 13000, 17000 and 14700. Find arithmetic mean.

    [Ans: Mean = 14650]

CHAPTER 3

Measures of Dispersion

3.1. MEASURES OF DISPERSION

3.1.1. Meaning and Definition

Distribution cannot be clearly depicted by measuring the averages or central tendency. Averages provide the observations of only the central part of the distribution. So, study of scatteredness of observations is very important and this study is known as **measure of dispersion**. The word 'Dispersion' literally means 'fluctuation', 'scatter', 'variation', 'deviation', or 'spread'. So, measure of dispersion shows the variation of an individual item from its average. It is also known as **measure of variation** as it explains the extent of scatterness in an observation and it measures the mean deviation about some central value.

Dispersion is based on values from first order (the average or mean). So, it is known as **second order**. Deviation is calculated by measuring the difference between mean and actual value.

According to A.L. Bowley, "Dispersion is the measure of the variation of the items."

According to Reigleman, "Dispersion is the extent to which the magnitudes or qualities of the items differ; that is the degree of diversity".

According to Connor, "Dispersion is a measure of the extent to which the individual vary."

According to Kafka, "The measurement of a scatteredness of the mass of figures in a series about an average is called measure of dispersion or measure of variation".

3.1.2. Methods of Measuring Dispersion

The two types of methods for measuring dispersion are as follows:

1) **Absolute Measures of Dispersion:** The measure of dispersion which is expressed in terms of the units of the observations (e.g., Rupees, Meter, Years, etc.) is called absolute measure of dispersion.
   i) Range,
   ii) Quartile deviation,
   iii) Mean deviation, and
   iv) Standard deviation.

2) **Relative Measures of Dispersion:** The measure of dispersion which is independent of unit or may involve the point about which the deviations are taken, is known as relative measure of dispersion or coefficient of dispersion. There is only one relative measure corresponding to an absolute measure of dispersion.
   i) Coefficient of range,
   ii) Coefficient of quartile deviation,
   iii) Coefficient of mean deviation, and
   iv) Coefficient of variance.

Fundamentals of range and standard deviation have been discussed in detail further in the chapter.

Scanned with CamScanner

3.2. RANGE

3.2.1. Introduction

Range is the simplest absolute measure of dispersion which shows the difference between the highest and the lowest value in a series.

Mathematically: $R = L - S$

Where,

R = Range,
L = Maximum (largest) value,
S = Minimum (smallest) value.

3.2.2. Advantages of Range

1) It is very easy to calculate and simple to understand.
2) It is rigidly defined.
3) It provides the limit (which is measured from set of data) within which all the data items occur.

3.2.3. Disadvantages of Range

1) This method is highly affected by the extreme items as it gives importance to the two extreme values.
2) This method does not provide any information about the structure of the series.
3) It is affected by the sampling fluctuations.
4) Range is not a reliable method to measure the dispersion.

3.2.4. Applications of Range

Applications/uses of range are as follows:

1) **In Quality Control:** Control charts are prepared for checking the qualities of a product. In this condition, range plays a very important role. In case, range increases beyond a specific point then the checking of the production machine is started.
2) **In Share Market:** Range is useful in the study of variations in share and stocks.
3) **In Weather Forecasting:** The difference between the maximum and minimum temperature is generally found by the meterological department. This information is useful to the general public.

3.2.5. Coefficient of Range

Relative measure of variation (coefficient of range) is used to compare the series. Coefficient of range is defined as:

$$\text{Coefficient of Range (C.R.)} = \frac{\text{Largest Value} - \text{Smallest Value}}{\text{Largest Value} + \text{Smallest Value}} = \frac{L-S}{L+S}$$

3.2.6. Methods of Calculation of Range

1) Calculation of Range - Individual Series
2) Calculation of Range - Discrete Series (Ungrouped Data)
3) Calculation of Range - Grouped Series (Grouped Data)

3.2.6.1. Calculation of Range - Individual Series

To find out the value of range in the individual series, one needs to calculate the difference between highest and lowest value of the series.

Example 1: Calculate Range and its Coefficient from the following data:
54, 45, 16, 15, 75, 85, 28

Solution: Here, Largest value (L) = 85, Smallest value (S) = 15
Range = L – S = 85 – 15 = 70

$$\text{Coefficient of Range} = \frac{L-S}{L+S} = \frac{85-15}{85+15} = \frac{70}{100} = 0.7 \text{ or } 70\%$$

3.2.6.2. Calculation of Range - Discrete Series (Ungrouped Data)

In case of individual series, it can also be calculated by taking the difference of highest and lowest value. In this case frequency of the series is not taken into consideration for the calculation.

Example 2: Find range and coefficient of range from the following data:

X	5	15	25	35	45	55
f	7	14	18	10	4	1

Solution: Here, Largest value (L) = 55, Smallest value (S) = 5

$\therefore$ Range = L – S = 55 – 5 = 50

$$\text{Coefficient of Range} = \frac{L-S}{L+S} = \frac{55-5}{55+5} = \frac{50}{60} = 0.83 \text{ or } 83$$

3.2.6.3. Calculation of Range - Grouped Series (Grouped Data)

In case of grouped series it can be calculated by taking the difference of lower limit of the lowest class interval and upper limit of the highest class interval. It can also be measured by taking the difference of mid value of the lowest class interval and mid value of the highest class interval.

Example 3: Calculate Range and the Coefficient of range from the following data:

X	10 – 20	20 – 30	30 – 40	40 – 50	50 – 60	60 – 70	70 – 80	80 – 90	90 – 100
f	6	4	15	24	11	3	10	16	20

Solution: Largest value (L) = 100; Smallest value (S) = 10
$\therefore$ Range (L – S) = 100 – 10 = 90

$$\text{Coefficient of Range} = \frac{L-S}{L+S} = \frac{(100-10)}{(100+10)} = \frac{90}{110} = 0.82 \text{ or } 82\%$$

3.3. STANDARD DEVIATION (S.D.)

3.3.1. Introduction

Standard deviation is used to measure the spread of items in a set of observation. If all the observation values are identical or distribution of items of a set is uniform, then deviation of every value from mean is zero. When elements of the set are more dispersed, then standard deviation becomes larger. It is denoted by the small Greek letter σ (Sigma). If $\bar{X}$ is the mean of $X_1, X_2, \text{----}, X_N$, then,

$$\sigma = \sqrt{\left[\frac{1}{N}\left\{(X_1 - \bar{X})^2 + \text{.....} + (X_N - \bar{X})^2\right\}\right]} = \sqrt{\sum_{i=1}^{N} \frac{(X_i - \bar{X})^2}{N}}$$

Scanned with CamScanner

According to Yule and Kenndall, "The standard deviation is the square-root of the arithmetic mean of the square of all the deviations, deviations being measured from the arithmetic mean of the observation."

3.3.2. Properties of Standard Deviation

1) Standard deviation is always positive.
2) Standard deviation will not change if all the variables are summated with the same number.
3) If researcher multiplies the variable with the same number, then the output received will be multiplied by the square of that same number.
4) In case of multiple distribution (with same mean and known standard deviation), SD can be calculated as:
   i) if samples size are same

$$\sigma = \sqrt{\frac{\sigma_1^2 + \sigma_2^2 + ... + \sigma_n^2}{N}}$$

   ii) if samples size are different

$$\sigma = \sqrt{\frac{k_1.\sigma_1^2 + k_2.\sigma_2^2 + ... + k_N.\sigma_N^2}{k_1 + k_1 + ... + k_N}}$$

3.3.3. Advantages of Standard Deviation

1) S.D. is rigidly defined and considers all the observations for calculation.
2) It contains various mathematical properties and is used in algberic calculation.
3) Combined S.D. can be calculated for two or more groups.
4) Compared to other measures of dispersion, standard deviation is less affected by sampling fluctuation.
5) It is used to calculate coefficient of variation.

3.3.4. Disadvantages of Standard Deviation

1) More complex to calculate and understand, compared to other measures of dispersion.
2) Distance from the mean decides the weightage of the element. If distance (from the mean) is more, then it is given more weight and if distance is less (from mean), then that element is given less weight.

3.3.5. Methods of Calculation of Standard Deviation

1) Calculation of Standard Deviation – Individual Series
2) Calculation of Standard Deviation – Discrete series
3) Calculation of Standard Deviation – Continuous Series

3.3.5.1. Calculation of Standard Deviation – Individual Series

There are two methods of calculating standard deviation in an individual observation or series:

1) **Deviation Taken from Actual Mean:** This method is adopted when the mean is a whole number. The following are the steps:
   i) Find out the actual mean of the series.
   ii) Find out the deviation of each value from the mean (d_x or $x = X - \bar{X}$).
   iii) Square the deviations and take the total of squared deviations Σx^2.

   iv) Divide the total (Σx^2) by the number of observations. The square root of the quotient is standard deviation. Thus apply the following formula:

$$\sigma = \sqrt{\frac{\Sigma x^2}{N}} \text{ or } \sqrt{\frac{\Sigma (X - \bar{X})^2}{N}}$$

2) **Deviation Taken from Assumed Mean (Short-cut Method):** This method is adopted when the arithmetic average is a fractional value. Taking deviations from fractional value would be a very difficult and tedious task. To save time and labour, we apply short cut method; deviations are taken from an assumed mean. The following are the steps:
   i) Assume any one of the item in the series as an average (A).
   ii) Find out the deviations from the assumed mean; i.e., (X – A) denoted by d_x.
   iii) Find out the total of the deviations; i.e., Σd_x.
   iv) Square the deviations; i.e., d_x^2 and add up the squares of deviations, i.e., Σd_x^2.
   v) Then substitute the values in the following formula:

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

   Where,

   d_x stands for the deviation from assumed mean = (X – A).

Example 4: Find S.D of (₹) 8, 10, 15, 24, 28.

Solution: Let the assumed mean (A) = 16.

Calculation from Arithmetic Mean

X	$x = (X - \bar{X})$	x^2
8	–9	81
10	–7	49
15	–2	4
24	7	49
28	11	121
ΣX = 85		$\Sigma x^2 = 304$

Calculation from Assumed Mean

X	$d_x = (X - A)$	d_x^2
8	–8	64
10	–6	36
15	–1	1
24	8	64
28	12	144
	5	$\Sigma d_x^2 = 309$

From Arithmetic Mean

$$\bar{X} = \frac{\Sigma X}{N} = \frac{85}{5} = 17$$

$$\sigma = \sqrt{\frac{\Sigma x^2}{N}} \text{ or } \frac{\Sigma (X - \bar{X})^2}{N} = \sqrt{\frac{1}{5} \times 304} = \sqrt{60.8} = 7.8$$

From Assumed Mean

Let A (assumed mean) = 16

$$\sigma = \sqrt{\frac{\Sigma d_x^2}{N} - \left(\frac{\Sigma d_x}{N}\right)^2} = \sqrt{\frac{309}{5} - \left(\frac{5}{5}\right)^2} = \sqrt{61.8 - (1)^2} = \sqrt{60.8} = 7.8$$

Note: If the actual mean is in fraction, then it is better to take deviations from an assumed mean for avoiding too much calculation.

3.3.5.2. Calculation of Standard Deviation – Discrete Series (Ungrouped Data)

There are three methods for calculating standard deviation in discrete series:

1) **Actual Mean Method:** It includes the following steps:
   i) Calculate the mean of the series.

Scanned with CamScanner

ii) Find deviations for various items from the mean i.e., $(X-\bar{X}) = d_x$.

iii) Square the deviations d_x^2 and multiply by the respective frequencies (f). We get fd_x^2.

iv) Total the product (Σfd_x^2). Then apply the formula $\sigma = \sqrt{\frac{\Sigma f d_x^2}{\Sigma f}}$

Example 5: Calculate standard deviation from the following series:

Age (in years)	15	25	35	45	55	65
No. of Persons	7	25	20	16	11	6

Solution: Calculation of Standard Deviations

Age (X)	No. of Persons (f)	fX	$d_x = (X-\bar{X})$	fd_x^2
15	7	105	– 22	3388
25	25	625	– 12	3600
35	20	700	– 2	80
45	16	720	8	1024
55	11	605	18	3564
65	6	390	28	4704
	$\Sigma f = 85$	$\Sigma fX = 3145$		$\Sigma fd_x^2 = 16360$

$$\bar{X} = \frac{\Sigma fX}{\Sigma f} = \frac{3145}{85} = 37 \text{ Hence } \sigma = \sqrt{\frac{\Sigma f(d_x)^2}{\Sigma f}} = \sqrt{\frac{16360}{85}} = \sqrt{192.5} = 13.87$$

2) **Assumed Mean Method:** We applied short-cut method in individual observation. Similarly we can apply it in discrete series also. Here deviations are taken, not from the actual mean, but from an assumed mean. It involves the following steps:

i) Assume any one of the items in the series as an average and this is called assumed average and denoted by A.

ii) Find out the deviations from assumed mean, i.e., (X – A), and denote it by d_x.

iii) Multiply these deviations by the respective frequencies and get the Σfd_x.

iv) Square the deviations (d_x^2).

v) Multiply the squared deviations (d_x^2) by the respective frequencies (f) and get Σfd_x^2.

vi) Substitute the values in the following formula:

$$\sigma = \sqrt{\frac{\Sigma f d_x^2}{\Sigma f} - \left(\frac{\Sigma f d_x}{\Sigma f}\right)^2} \quad \text{Where } d_x = X - A$$

Example 6: Find the Standard Deviation of the following series:

X	10	11	12	13	14	Total
f	4	16	22	14	6	62

Solution: Let the assumed mean (A) = 12.

Calculation of Standard Deviations

X	f	$d_x = (X-A)$	d_x^2	fd_x	fd_x^2
10	4	–2	4	– 8	16
11	16	–1	1	– 16	16
12	22	0	0	0	0
13	14	1	1	14	14
14	6	2	4	12	24
Total	$\Sigma f = 62$			$\Sigma fd_x = 2$	$\Sigma fd_x^2 = 70$

$$\sigma = \sqrt{\frac{\Sigma f(d_x)^2}{\Sigma f} - \left(\frac{\Sigma fd_x}{\Sigma f}\right)^2} = \sqrt{\frac{70}{62} - \left(\frac{2}{62}\right)^2}$$

$$= \sqrt{1.13 - 0.001} = 1.06$$

3) **Step Deviation Method:** Here we take a common factor for all the items of the series. In this method the calculation becomes easy and simple. The formula for this is:

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

Where, $d_x' = \frac{X-A}{i}$, i = Common Factor.

Example 7: Find the Standard Deviation for the following distribution:

X	4.5	14.5	24.5	34.5	44.5	54.5	64.5
f	5	3	7	18	14	9	4

Solution: Let the assumed mean (A) = 34.5.

Calculation of Standard Deviation

X	f	$d_x = (X-A)$	$d_x' = \left(\frac{d_x}{10}\right)$	fd_x'	$f(d_x')^2$
4.5	5	– 30	– 3	– 15	45
14.5	3	– 20	– 2	– 6	12
24.5	7	– 10	– 1	– 7	7
34.5	18	0	0	0	0
44.5	14	10	1	14	14
54.5	9	20	2	18	36
64.5	4	30	3	12	36
	$\Sigma f = 60$			$\Sigma fd_x' = 16$	$\Sigma f(d_x')^2 = 150$

$$\sigma = \sqrt{\left\{\frac{\Sigma f(d_x')^2}{\Sigma f} - \left(\frac{\Sigma fd_x'}{\Sigma f}\right)^2\right\}} \times i = \sqrt{\frac{150}{60} - \left(\frac{16}{60}\right)^2} \times 10$$

$$= \sqrt{2.5 - 0.071} \times 10 = 1.56 \times 10 = 15.6$$

3.3.5.3. Calculation of Standard Deviation – Continuous Series (Grouped Data)

In the continuous series the method of calculating standard deviation is almost the same as in a discrete frequency distribution. But in a continuous series, mid-values of the class intervals are to be found out. The step deviation method is widely used.

It involves the following steps:

1) Find out the mid-value of each group or class.
2) Assume one of the mid-values as an average and denote it by A.
3) Find out deviation of each mid-value from the assumed average A and denote these deviations by d_x.
4) If the class-intervals are equal, then take a common factor. Divide each deviation by the common factor and denote this column by d_x'.

Scanned with CamScanner

5) Multiply these deviations d_x' by the respective frequencies and get $\Sigma fd_x'$.
6) Square the deviations and get $d_x'^2$.
7) Multiply the squared deviation ($d_x'^2$) by the respective frequencies (f). Then obtain the total: $\Sigma fd_x'^2$.
8) Substitute the values in the following formula to get the standard deviation.

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

Where, $d_x' = \frac{X - A}{i}$, i = Common factor

Example 8: Find the S.D. from the following figures:

Height (Inches)	44-46	46-48	48-50	50-52	52-54	Total
No. of Children	5	25	28	22	5	85

Solution: Let A (assumed mean) = 49

Height (Inches)	Mid-point (X)	No. of Children (f)	$d_x = X - 49$	$d_x' = d/2$	fd_x'	$fd_x'^2$
44-46	45	5	– 4	– 2	– 10	20
46-48	47	25	– 2	– 1	– 25	25
48-50	49	28	0	0	0	0
50-52	51	22	2	1	22	22
52-54	53	5	4	2	10	20
Total		$\Sigma f = 85$			$\Sigma fd_x' = -3$	$\Sigma fd_x'^2 = 87$

$$\sigma = \sqrt{\left\{\frac{\Sigma f(d_x')^2}{\Sigma f} - \left(\frac{\Sigma d_x'}{\Sigma f}\right)^2\right\}} \times i = \sqrt{\frac{87}{85} - \left(\frac{-3}{85}\right)^2} \times 2$$

$$= \sqrt{\left(\frac{7395 - 9}{7225}\right)} \times 2 = \sqrt{\frac{7386}{7225}} \times 2 = 1.01 \times 2 = 2.02$$

3.4. PHARMACEUTICAL PROBLEMS

Example 9: Calculate Range and its Coefficient from the following data:

54, 45, 16, 15, 75, 85, 28

Solution: Here, Largest value (L) = 85, Smallest value (S) = 15

Range = L – S = 85 – 15 = 70

$$\text{Coefficient of Range} = \frac{L - S}{L + S} = \frac{85 - 15}{85 + 15} = \frac{70}{100} = 0.7 \text{ or } 70\%$$

Example 10: Find range and coefficient of range from the following data:

X	5	15	25	35	45	55
f	7	14	18	10	4	1

Solution: Here, Largest value (L) = 55, Smallest value (S) = 5

∴ Range = L – S = 55 – 5 = 50

$$\text{Coefficient of Range} = \frac{L - S}{L + S} = \frac{55 - 5}{55 + 5} = \frac{50}{60} = 0.83 \text{ or } 83\%$$

Example 11: Find the standard deviation of 60, 85, 45, 74, 56, 42, 98 and 76.

Solution:

Calculation of Standard Deviation

X	$(X - \bar{X})$	$(X - \bar{X})^2$
60	– 7	49
85	18	324
45	– 22	484
74	7	49
56	– 11	121
42	– 25	625
98	31	961
76	9	81
$\Sigma X = 536$		$\Sigma(X - \bar{X})^2 = 2694$

$$\text{Mean } (\bar{X}) = \frac{\Sigma X}{N} = \frac{536}{8} = 67$$

$$\sigma = \sqrt{\frac{\Sigma(X - \bar{X})^2}{N}} = \sqrt{\frac{2694}{8}} = \sqrt{336.75} = 18.35$$

Example 12: Calculate the standard deviation from the following data:

Marks in Economics	10	20	30	40	50	60	70	80	90
No. of Students	2	6	9	12	15	29	11	9	7

Solution: Calculation of Standard Deviation from Actual Mean

Marks in Economics (X)	No. of Students (f)	fX	$d_x = (X - \bar{X})$	fd_x^2
10	2	20	– 45	4050
20	6	120	– 35	7350
30	9	270	– 25	5625
40	12	480	– 15	2700
50	15	750	– 5	375
60	29	1740	5	725
70	11	770	15	2475
80	9	720	25	5625
90	7	630	35	8575
	$\Sigma f = 100$	$\Sigma fX = 5500$		$\Sigma fd_x^2 = 37500$

$$\bar{X} = \frac{\Sigma fX}{\Sigma f} = \frac{5500}{100} = 55$$

$$\text{Standard Deviation } (\sigma) = \sqrt{\frac{\Sigma fd_x^2}{\Sigma f}}$$

$$= \sqrt{\frac{37500}{100}} = \sqrt{375} = 19.36$$

Scanned with CamScanner

Example 13: Sample of the weight (kg) of 86 medical students are given in table below:

Table 3.1: Raw Data - Weights (kg) of a Sample of 86 Medical Student

Lowest Value	66.32	74.23	79.12	83.76	88.24	90.01	98.54
42.34	66.56	74.34	79.43	84.32	88.43	90.43	98.54
51.56	67.33	75.32	79.76	84.87	88.54	91.23	99.35
53.54	68.92	75.43	80.03	85.33	88.65	92.46	99.75
58.69	69.12	75.78	81.23	85.55	88.65	94.56	100.54
60.32	70.33	76.78	81.24	85.63	88.67	95.43	104.23
60.94	71.23	77.65	81.34	85.78	88.75	95.45	106.45
61.44	71.28	77.67	82.34	85.78	89.46	96.45	107.35
62.55	72.75	77.96	82.43	86.43	89.55	96.54	107.52
64.32	73.43	78.45	83.45	87.54	89.64	97.45	109.35
65.87	73.65	78.54	**83.45**	87.56	89.89	97.46	**Highest Value**

Determine the range and standard deviation.

Solution: The ranges of the weights of medical students are from 42.34 to 109.35kg. The variance is:

$$\sigma^2 = \frac{(42.34-82.303)^2+(51.56-82.303)^2+\ldots+(109.35-82.303)^2}{86-1} = \frac{15252123}{85} = 179.437 kg^2$$

The standard deviation is $\sigma = \sqrt{179.437} = 13.395 kg$

Example 14: A batch of bolus tablets is awaiting release from the quality control laboratories of pharmaceutical company. One of the features of the bolus tablets that must be quantified is the time required for the release of 50% of the original drug loading of the bolus ($t_{50\%}$) using the dissolution test. **Table 3.2** illustrates the $t_{50\%}$ of 15 boluses randomly sampled from the batch. Find the mean and standard deviation of the time needed for the release of 50% of the original drug loading of the bolus.

Table 3.2: Time required for Dissolution of 50% of the Original Mass of Drug from 15 Bolus Tablets

Bolus Number	$T_{50\%}$(h)
1	20.2
2	21.6
3	24.5
4	28.6
5	22.6
6	24.0
7	21.9
8	22.0
9	26.1
10	25.3
11	23.4
12	20.1
13	23.5
14	24.0
15	25.9

Solution: Calculation the Mean

$$\bar{X} = \frac{\left(20.2+21.6+24.5+28.6+22.6+24.0+21.9+22.0+26.1+25.3+23.4+20.1+23.5+24.0+25.9\right)}{15} = \frac{353.7}{15} = 23.6h$$

Calculation of Sample Standard Deviation

$$\sigma = \sqrt{\frac{\Sigma X^2 - \left\{\frac{(\Sigma X)^2}{N}\right\}}{N-1}}$$

X	X^2
20.2	408.0
21.6	466.6
24.5	600.3
28.6	818.0
22.6	510.8
24.0	576.0
21.9	479.6
22.0	484.0
26.1	681.2
25.3	640.1
23.4	547.6
20.1	404.0
23.5	552.3
24.0	576.0
25.9	670.8
ΣX = 53.7h	ΣX^2 = 8415.3h^2

Thus:

$$\sigma = \sqrt{\frac{\Sigma X^2 - \left\{\frac{(\Sigma X)^2/N}{N}\right\}}{N-1}} = \sqrt{\frac{8415.3 - \left\{\frac{(353.7)^2/15}{15}\right\}}{14}} = 2.3h$$

Thus, the mean and standard deviation of the $t_{50\%}$ of the sample removed from a batch is 23.6 ± 2.3h.

Example 15: Using an iodometric technique, the concentrations (mg/5mL) of a penicillin antibiotic have been examined in five separate bottles of a paediatric suspension. Find the mean and standard deviation. Also determine the consideration of the contribution of each observation to the sample variance.

Solution: Calculation the Mean

$$\bar{X} = \frac{(125+124+121+123+16)}{5} = 101.8 mg/5mL$$

Calculation of the Variance

$$\sigma^2 = \frac{\Sigma(X_i-\bar{X})^2}{N-1} = \left[\frac{(125.0-101.8)2+(124.0-101.8)^2+\ldots..+(16.0-101.8)^2}{4}\right]$$

$$= 2302.7 (mg/5mL)^2$$

Since the square root of the variance is standard deviation, hence it may be easily calculated as follows:

$$\sigma = \sqrt{2302.7} = 48.0 mg/5mL$$

Scanned with CamScanner

Consideration of Contribution of Each Observation to the Sample Variance

The contribution of each observation in overall variance (five bottles) and also the standard deviation calculated by using standard formula:

$$\sigma^2 = \frac{\Sigma(X_i - \bar{X})^2}{N-1}$$

For example, the recorded concentration of penicillin in bottle 1 was 125mg/5mL. Thus, we have:

$$\sigma_1^2 = \frac{(125-101.8)^2}{4} = 134.56(mg/5mL)^2$$

The data are shown in table 3.3:

Table 3.3: Individual Contributions of Every Vial to the Total Variance

Bottle Number	Concentration of Penicillin (mg/5mL)	Contribution to the Overall Variance
1	125	134.5
2	124	123.2
3	121	92.2
4	123	112.4
5	16	1840.4
		$s^2 = 2302.7 (mg/5mL)^2$

In this example, concentration of penicillin of five bottles is represented by outlier and also indicates a problem with either the manufactured batch or the analytical method. As it is known that the distributions in which the mean is numerically greater than the median are positively skewed. This data describe such a distribution where Means = 132.4mg/5mL and Median 123mg/5mL.

3.5. EXERCISE

3.5.1. Very Short Answer Type Questions

1) What is the meaning of dispersion?
2) Discuss about the measures of dispersion.
3) Define the term Range.
4) What is coefficient of range?

3.5.2. Short Answer Type Questions

1) Explain the method of calculation of range with example.
2) What is standard deviation?
3) Write the advantages and disadvantages of standard deviation.
4) Write the properties of standard deviation.
5) Explain the methods of standard deviation and also give an example.

3.5.3. Long Answer Type Questions

1) Calculate Range and its Coefficient from the following data:
   53, 46, 18, 16, 75, 84, 28

   [Ans: Coefficient of range = 0.65]

2) Compute for the following data:
   105, 29, 19, 109, 26, 40, 111, 21, 105, 77, 73, 108, 66, 95, 25
   i) Range
   ii) Coefficient of Range

   [Ans: i) Range = 92 ii) Coefficient of Range = 0.7]

3) Calculate the standard deviation for the following distribution:

Height (Inches)	No. of Persons	Height (Inches)	No. of Persons
60 less than 63	4	69 less than 72	33
63 less than 66	14	72 less than 75	8
66 less than 69	59	75 less than 78	2

   [Ans: σ=2.79]

4) Blood serum cholesterol levels of 10 persons are as under:
   240, 260, 290, 245, 255, 288, 272, 263, 277, 250. Calculate standard deviation with the help of assumed mean.

   [Ans: σ =16.48]

5) Calculate standard deviation from the following data:

Age in years	4 – 6	6 – 8	8 –10	10 – 12	12 – 14	14 – 16	16 – 18
No. of students	30	90	120	150	80	60	20

   [Ans: σ=3]

6) Calculate the standard deviation for the following distribution:

Class	30 – 40	40 – 50	50 – 60	60 – 70	70 – 80	80 – 90	90 –100
Frequency	3	7	12	15	8	3	2

   [Ans: S.D = 14.18]

7) Calculate S.D. from the following:

Marks (below)	20	40	60	80	100
No. of Students	8	20	50	70	80

   [Ans: σ =22.8]

Scanned with CamScanner

CHAPTER 4 Correlation

4.1. CORRELATION

4.1.1. Meaning and Definition

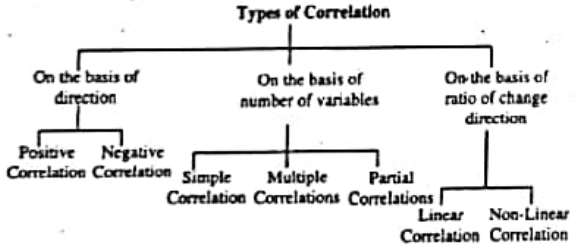
Study of the linear relationship between two variables is known as correlation. If there are two variables and changes in the value of one variable will affect the value of the other variable then both the variables are correlated. **Correlation** is a statistical tool used to measure the relationship between two sets of variables and express each in a precise manner.

According to Croxton and Cowden, "The appropriate statistical tool for discovering and measuring the relationship of quantitative nature and expressing it in brief formula is known as correlation".

According to A. M. Tuttle, "An analysis of the covariance of two or more variables is usually called correlation".

4.1.2. Types of Correlation

Categorisation of correlation is done on various bases such as on the direction of change of the variable, constancy of the ratio of change between the variable, etc.



4.1.2.1. Positive Correlation and Negative Correlation

Depending on the direction of change of the variables, correlation may be positive or negative.

1) **Positive Correlation:** Correlation is positive when two variables vary in the same direction. **For example,** consider the correlation between sales and expenses of a firm **(table below).**

Firms	1	2	3	4	5	6
Sales (in ₹)	50	60	70	80	90	100
Expenses (in ₹)	11	15	19	23	27	31

2) **Negative Correlation:** Two variables are said to be negatively correlated when both the variables vary in the opposite direction. When one variable increases then other variable decreases and *vice versa* is also a negative correlation. **For example,** consider the correlation between production and price of crop.

Production (in kg)	100	200	300	400	500	600	700	800
Price (₹ Per kg)	10	9	8	7	6	5	4	3

4.1.2.2. Linear Correlation & Non-Linear (Curvilinear) Correlation

On the basis of constancy of the ratio of change between the variable, correlation may be linear or non-linear.

1) **Linear Correlation:** In a linear correlation, change in the values of one variable has a fixed ratio to the variation in the values of other variable. When these variables are plotted on a graph, then plotted points would fall on a straight line. **For example,** consider the following relationship shown in the table.

Mustard (kg)	10	20	30	40	50	60	70	80
Oil (Kg)	3	6	9	12	15	18	21	24

2) **Non-Linear (Curvilinear) Correlation:** In a non-linear correlation, change in values of one variable does not have a fixed ratio to the variation in the value of other variable. When these variables are plotted on a graph, then plotted points would fall on a curve. **For example,** consider the following relationship shown in the table.

Use of Fertilizer (in kg.)	1	2	3	4	5
Production of Rice (in kg.)	4	6	9	17	28

4.1.2.3. Simple, Partial and Multiple Correlation

Depending on the number of variables studied, it can be categorised into three types:

1) **Simple Correlation:** When we measure the linear relationship between two variables then this interpretation is known as simple correlation, **e.g.,** relationship between sales and expenses, income and consumption etc.

2) **Partial Correlation:** If we have various related variables and try to find out the relationship between two variables then it is known as partial correlation. **For example,** consider the two variables height and weight, which are partially correlated because of the effect of third variable 'age' on height and weight. In this condition, if we neglect the effect of age, and study the relationship between height and weight, then this correlation is known as partial correlation.

3) **Multiple Correlation:** It is defined as the measurement of the effect of multiple variables on one variable. **For example,** if we try out to find the relationship of rainfall and temperature on the yield of wheat, then this is known as multiple correlation.

4.1.3. Degree of Correlation

On the basis of coefficient of correlation, the degree of correlation is determined as follows:

1) **Perfect Correlation:** If two variables change in the same direction and in the same proportion, the correlation between the two variables is called perfect positive correlation. According to Karl Pearson the coefficient of correlation in this case is +1. If two variables change in the opposite direction and in the same proportion, the correlation between the two variables is called **perfect negative correlation.** In this case the coefficient of correlation is – 1.

2) **Absence of Correlation:** If two series of two variables show no relation between them or change in one variable does not lead to a change in the other variable, then it means that there is no relationship between variables. In this case, the coefficient of correlation is zero.

3) **Limited Degree of Correlation:** If two variables are not perfectly correlated or there is an absence of perfect correlation, then it is referred as Limited correlation. It may be positive, negative or zero but lies with the limits ±1. High degree, moderate degree and low degree are the three categories of this kind of correlation. If the value of coefficient of correlation (represented by 'r') is lies between ±0.75 to ±1, it is known as **High degree** of correlation and when the value of r lies between ±0.25 to

Scanned with CamScanner

Table 4.1: Degree of Correlation

Degree	Positive Correlation Coefficient	Negative Correlation Coefficient
1. Perfect	+1	-1
2. Limited		
a) High	Between +0.75 to +1	Between -0.75 to -1
b) Moderate	Between +0.25 to +0.75	Between -0.25 to -0.75
c) Low	Between 0 to 0.25	Between 0 to -0.25
3. Absence	Zero	Zero

4.1.4. Correlation Coefficient

The extent of measurement as to how much one number can expect to be influenced by changes in another number is known as correlation coefficient. **For example,** measuring the value that how much the share price of a company is expected to get influenced, when there are fluctuations in the index. If the two numbers are perfectly correlated then it means that value of correlation coefficient is +1.

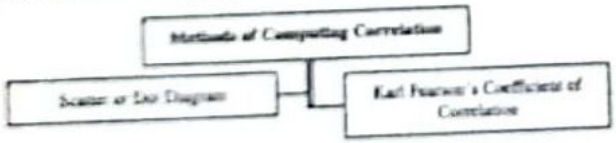
If one grows the other also grows, and the change in one is a multiple of the change in the other. If the numbers are perfectly inversely correlated then it means that value of correlation coefficient is –1. If one grows the other falls. The growth in one variable is a negative multiple of the growth in the other variable. If the two variables are not related then the correlation coefficient is zero.

When numbers are related and the coefficient is either 1 or -1 and other influences and relationship between two numbers is not fixed then this is known as non-zero correlation coefficient. If one number is known, the other number can be calculated, but with uncertainty. Uncertainty increases when correlation coefficient is more towards to zero. But this relationship is not useful, if the correlation coefficient is very low.

4.2. METHODS OF COMPUTING CORRELATION

4.2.1. Introduction

The various method of computing linear correlation can be shown diagrammatically as follows:



4.2.2. Scatter or Dot Diagram

Graphical presentation of the relationship between two variables calculated on the same set of individual is known as **scatter diagram. For example,** if we plot a dot chart for some predefined value of two variables and get some trends (either upwards or downwards) in the chart, then the trend shows a correlation between both the variables. The idea will be clear from the **figure 4.1**

1) In **figure 4.1(a)**, the values of the two variables move in the upward direction, hence the correlation is positive.
2) In **figure 4.1 (b)**, we find negative correlation as the values move in reverse direction.

3) In **figures 4.1 (c)**, we don't get any trend lying and hence it shows the absence of correlation and $r = 0$.

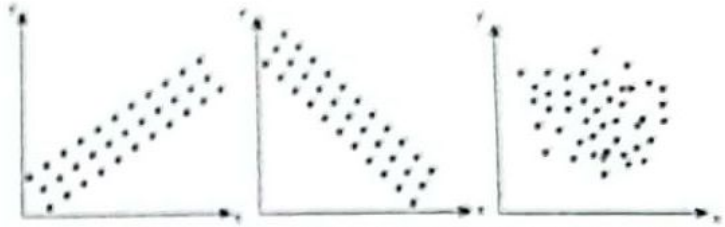


Figure 4.1

4) In **figure 4.2**, the value of the two variables moves in the same direction and when $r = +1$ then its shows perfect positive correlation and when $r = -1$ then its shows perfect negative correlation.

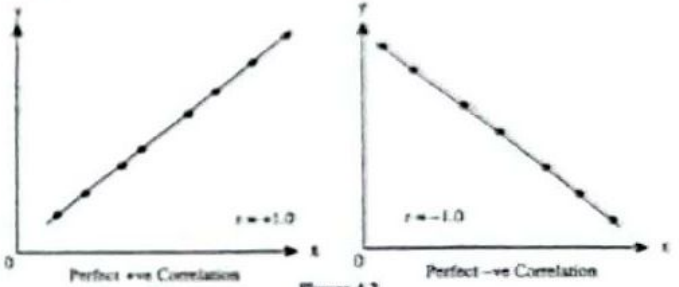


Figure 4.2

Example: When human take the various tablets then it can be demonstrated a correlation between time to 80% dissolution of various tablet formulations of a particular drug and the fraction of the dose absorbed. **Figure 4.3** shows that as dissolution increases (i.e., the time to 80% dissolution decreases) in vivo absorption increases. Scatter plots consists of data pairs, X and Y, both are variable. In this example, dissolution time and fraction absorbed are both taken as random variables.

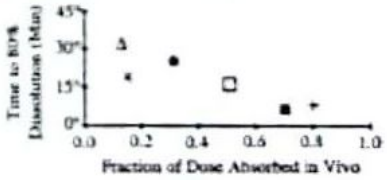


Figure 4.3: Scatter Plot Showing the Correlation of Dissolution Time and in vivo Absorption of Six Tablet Formulation. Δ, Formulation A; ×, Formulation B; ·, Formulation C; □, Formulation D; ■, Formulation E; +, Formulation F.

4.2.2.1. Advantages of Scatter Diagram

1) It is a very simple and non-mathematical technique.
2) It is not influenced by the size of extreme items.
3) It is the very basic step to find out the relationship between two variables.

Scanned with CamScanner

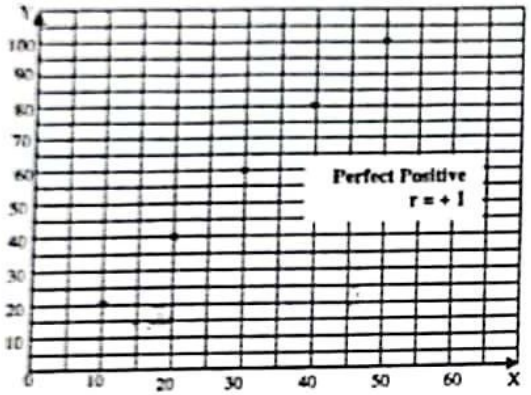
4.2.2.2. Disadvantages of Scatter Diagram

The main disadvantage of this technique is that it cannot find out an exact degree of correlation between two variables. We can only view the visual form of correlation and direction (positive or negative) on the chart.

Example 1: Draw a scatter diagram for the following and indicate whether the correlation is positive or negative.

X	10	20	30	40	50
Y	20	40	60	80	100

Solution:



4.2.3. Karl Pearson's Coefficient of Correlation

To calculate the magnitude of linear relationship between two variables Karl Pearson (biometrician and statistician) gave a quantitative technique. This technique is known as **Pearsonian Coefficient of Correlation** (r) and is extensively used in practice.

4.2.3.1. Properties of Karl Pearson's Coefficient of Correlation

1) Ideal measure of correlation and is independent of the units of the variables.
2) Free from change of origin and scale.
3) Based on all observations.
4) Lies between −1 and +1 and when:
   i) $r = -1$, shows a perfect negative correlation between variables.
   ii) $r = 0$, then there is no correlation between variables.
   iii) $r = +1$, shows a perfect positive correlation between variables.

4.2.3.2. Advantages of Karl Pearson's Coefficient of Correlation

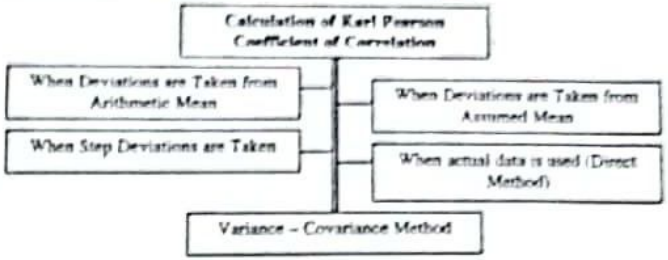
The main advantage of the Karl Pearson's coefficient of correlation is that it gives the result in one value and also summarizes the degree of correlation and direction.

4.2.3.3. Disadvantages of Karl Pearson's Coefficient of Correlation

1) Every time assumes only a linear relationship between variables.
2) Interpreting the value of correlation coefficient (r) is difficult.
3) Significance of Correlation coefficient (r) is affected by the extreme values.
4) It is a time consuming method.
5) Does not convey the cause and effect relationship.

4.2.3.4. Calculation of Karl Pearson Coefficient of Correlation

The various formulae used to calculate co-efficient of (r) are:



4.2.3.5. When Deviations are Taken from Arithmetic Mean

When deviation is taken from arithmetic mean, then r is calculated as:

$$r = \frac{\sum xy}{\sqrt{\sum x^2}\sqrt{\sum y^2}} \text{ or } \frac{\sum xy}{\sqrt{\sum x^2 \cdot \sum y^2}} \text{ or } \frac{\sum xy}{N\sigma_x \sigma_y} \text{ Or } r = \frac{\sum (X-\bar{X})(Y-\bar{Y})}{\sqrt{\sum (X-\bar{X})^2 \sum (Y-\bar{Y})^2}}$$

Where, $x = X - \bar{X}$, (deviation from Arithmetic Mean (A.M.) of X series)

$y = Y - \bar{Y}$, (deviation from Arithmetic Mean (A.M.) of Y series)

σ_x = Standard Deviation (S.D.) of X series

σ_y = Standard Deviation (S.D.) of Y series

And N = Number of observation.

Example 2: Calculate Karl Pearson Coefficient of Correlation.

X	42	52	55	60	66	68	65	60	58	34
Y	11	13	18	22	26	40	31	27	24	18

Solution:

X	Y	$x = (X - \bar{X})$	x^2	$y = (Y - \bar{Y})$	y^2	xy
42	11	− 14	196	− 12	144	168
52	13	− 4	16	− 10	100	40
55	18	− 1	1	− 5	25	5
60	22	4	16	− 1	1	− 4
66	26	10	100	3	9	30
68	40	12	144	17	289	204
65	31	9	81	8	64	72
60	27	4	16	4	16	16
58	24	2	4	1	1	2
34	18	− 22	484	− 5	25	110
$\Sigma X = 560$	$\Sigma Y = 230$		$\Sigma x^2 = 1058$		$\Sigma y^2 = 874$	$\Sigma xy = 643$

$$\bar{X} = \frac{\sum X}{N} = \frac{560}{10} = 56; \quad \bar{Y} = \frac{\sum Y}{N} = \frac{230}{10} = 23$$

Coefficient of correlation (r) $= \frac{\sum xy}{\sqrt{\sum x^2 \times \sum y^2}} = \frac{643}{\sqrt{1058 \times 643}}$

$$= \frac{643}{\sqrt{680294}} = \frac{643}{824.8} = 0.779$$

Scanned with CamScanner

Example 3: Calculate Karl Pearson coefficient of correlation from the following data:

X	18	20	21	22	27	27	28	29	29	29
Y	23	37	29	28	28	31	35	30	36	33

Solution:

Calculation of Coefficient of Correlation

X	Y	$x=(X-\bar{X})$	$y=(Y-\bar{Y})$	x^2	y^2	xy
20	38	−7	−8	49	64	56
22	40	−5	6	25	36	−30
24	37	−4	−2	16	4	8
25	35	−3	−3	9	9	9
25	33	2	−3	4	9	−6
26	34	2	0	4	0	0
27	35	3	4	9	16	12
29	36	4	−1	16	1	−4
30	27	4	5	16	25	20
32	25	4	2	16	4	8
Σx = 250	Σy = 310	0		164	168	73

$$\bar{X}=\frac{\Sigma X}{N}=\frac{250}{10}=25 \quad \text{and} \quad \bar{Y}=\frac{\Sigma Y}{N}=\frac{310}{10}=31$$

Karl Pearson, coefficient of correlation

$$(r)=\frac{\Sigma xy}{\sqrt{\Sigma x^2\times\Sigma y^2}}=\frac{73}{\sqrt{164\times168}}=\frac{73}{\sqrt{27552}}=\frac{73}{165.99}=0.439$$

4.2.3.6. When Deviations are Taken from Assumed Mean

When deviations are taken from assumed mean, then the following formula is used to calculate the value of r.

$$r=\frac{N\times\Sigma d_x d_y-\Sigma d_x.\Sigma d_y}{\sqrt{N.\Sigma d_x^{\,2}-(\Sigma d_x)^2}\sqrt{N.\Sigma d_y^{\,2}-(\Sigma d_y)^2}}$$

Where $d_x=X-A_x$ and $d_y=Y-A_y$

A = Assumed Mean

Example 4: Find the coefficient of correlation from the following data

X	1	2	3	4	5	6	7	8	9	10
Y	62	56	48	41	36	28	21	16	12	8

Solution: Let the assumed mean for X is 5, Assumed mean for Y is 40.

X	Y	$d_x=(X-5)$	$d_y=(Y-40)$	d_x^2	d_y^2	d_xd_y
1	62	−4	22	16	484	−88
2	56	−3	16	9	256	−48
3	48	−2	8	4	64	−16
4	41	−1	1	1	1	−1
5	36	0	−4	0	16	0
6	28	1	−12	1	144	−12
7	21	2	−19	4	361	−38
8	16	3	−24	9	576	−72
9	12	4	−28	16	784	−112
10	8	5	−32	25	1024	−160
Total		$d_x=5$	$d_y=-72$	$\Sigma d_x^{\,2}=85$	$\Sigma d_y^{\,2}=3710$	$\Sigma d_xd_y=-547$

$$r=\frac{N\Sigma d_xd_y-\Sigma d_x\Sigma d_y}{\sqrt{\{N\Sigma d_x^{\,2}-(\Sigma d_x)^2\}\{N\Sigma d_y^{\,2}-(\Sigma d_y)^2\}}}$$

$$=\frac{10(-547)-5(-72)}{\sqrt{10\times85-(5)^2}\times\sqrt{10\times3710-(-72)^2}}$$

$$=\frac{-5470+360}{\sqrt{850-25}\times\sqrt{37100-5184}}=\frac{-5110}{\sqrt{825}\times\sqrt{31916}}=\frac{-5110}{\sqrt{26330700}}=\frac{-5110}{5131.34}$$

$r=-0.996$

The variable X and Y are perfectly negatively correlated.

4.2.3.7. When Step Deviations are Taken

To calculate the r following formula is used when step deviation is taken.

$$r=\frac{N\times\Sigma d'_xd'_y-\Sigma d'_x.\Sigma d'_y}{\sqrt{N.\Sigma d'^{\,2}_x-(\Sigma d'_x)^2}\sqrt{N.\Sigma d_y^{\,2}-(\Sigma d'_y)^2}};$$

Where $d'_x=\frac{d_x}{i_x}, d'_y=\frac{d_y}{i_y}$

Example 5: Calculate Coefficient of Correlation between the expenses and saving of any family.

Expenses (₹)	0	5	25	35	75
Saving (₹)	70	55	35	25	15

Solution: Let the assumed mean for X is 25, Assumed mean for Y is 40.

X	$d_x=(X-25)$	$d'_x=\frac{dx}{5}$	$(d'_x)^2$	Y	$d_y=(Y-30)$	$d_y'=\frac{dy}{5}$	$(d'_y)^2$	$d_x'd_y'$
0	−25	−5	25	70	25	5	25	−25
15	−10	−2	4	55	10	2	4	−4
25	0	0	0	35	−10	−2	4	0
35	10	2	4	25	−20	−4	16	−8
75	30	6	36	15	−30	−6	36	−36
N = 5		$\Sigma d_x'=1$	$\Sigma d_x'^2=69$	N = 5		$\Sigma d_y'=5$	$\Sigma d_y'^2=85$	$\Sigma d_x'd_y'=-73$

$$r=\frac{\Sigma d'_xd'_y-\frac{\Sigma d'_x\Sigma d'_y}{N}}{\sqrt{\Sigma d'^{\,2}_x-\frac{(\Sigma d'_x)^2}{N}}\sqrt{\Sigma(d'_y)^2-\frac{(\Sigma d'_y)^2}{N}}}$$

$$=\frac{-73-\left(-\frac{73}{5}\right)}{\sqrt{69-\frac{1}{5}}\sqrt{85-\frac{(-5)^2}{5}}}=\frac{-58.4}{\sqrt{68.8}\times\sqrt{80}}=\frac{-58.4}{\sqrt{5504}}=\frac{-58.4}{74.2}=-0.787$$

4.2.3.8. When Actual Data is Used (Direct Method)

In case of actual data following formula is used to calculate the coefficient r.

$$r=\frac{N\times\Sigma XY-\Sigma X.\Sigma Y}{\sqrt{N.\Sigma X^2-(\Sigma X)^2}\sqrt{N.\Sigma Y^2-(\Sigma Y)^2}}$$

Scanned with CamScanner

Example 6: Calculate the Karl Pearson coefficient of correlation for the following data:

X	4	7	11	14	19	15
Y	18	16	17	19	19	21

Solution:

X	Y	X^2	Y^2	XY
4	18	16	324	72
7	16	49	256	112
11	17	121	289	187
14	19	196	361	266
19	19	361	361	361
15	21	225	441	315
$\Sigma X = 70$	$\Sigma Y = 110$	$\Sigma X^2 = 968$	$\Sigma Y^2 = 2032$	$\Sigma XY = 1313$

Coefficient of Correlation,

$$r = \frac{N\sum XY - (\sum X)(\sum Y)}{\sqrt{N\sum X^2 - (\sum X)^2}\sqrt{N\sum Y^2 - (\sum Y)^2}} = \frac{6\times1313 - 70\times110}{\sqrt{6\times968 - (70)^2}\sqrt{6\times2032 - (110)^2}}$$

$$= \frac{7878 - 7700}{\sqrt{5808 - 4900}\sqrt{12192 - 12100}} = \frac{178}{\sqrt{908}\times\sqrt{92}} = \frac{178}{\sqrt{83536}} = \frac{178}{289} = 0.616$$

4.2.3.9. Variance – Covariance Method

Following formula is used to calculate the value of r when variance and covariance are given:

$$r = \frac{\text{Cov.}(X,Y)}{\sqrt{\text{Var.}(X)}\sqrt{\text{Var.}(Y)}}$$

Where, $r = \frac{\Sigma(X-\bar{X})(Y-\bar{Y})}{\sqrt{\Sigma(X-\bar{X})^2\,\Sigma(Y-\bar{Y})^2}}$

Divide Numerator and denominator by N we have,

$$r = \frac{\Sigma(X-\bar{X})(Y-\bar{Y})/N}{\sqrt{\frac{\Sigma(X-\bar{X})^2\,\Sigma(Y-\bar{Y})^2}{N}}} = \frac{\Sigma(X-\bar{X})(Y-\bar{Y})/N}{\sqrt{\frac{\Sigma(X-\bar{X})^2}{N}\times\frac{\Sigma(Y-\bar{Y})^2}{N}}} = \frac{\text{Cov.}(X,Y)}{\sqrt{\text{Var.}(X)}\sqrt{\text{Var.}(Y)}}$$

Example 7: If covariance between X and Y variables is 12.5 and the variance of X and Y are respectively 16.4 and 13.8, find the coefficient of correlation between them.

Solution: Covariance of XY = 12.5

Variance of X = 16.4; Variance of Y = 13.8

$$r = \frac{\text{Cov.}(X,Y)}{\sqrt{\text{Var.}(X)}\sqrt{\text{Var.}(Y)}} = \frac{12.5}{\sqrt{16.4}\sqrt{13.8}} = \frac{12.5}{\sqrt{226.32}} = \frac{12.5}{15.04} = 0.83 = 0.83$$

4.3. MULTIPLE CORRELATION

4.3.1. Introduction

The multiple correlation is a measure of linear relationship between a dependent variable and a number of independent variables. It is represented by R and can have any value between 0 and 1. For example, if one wants to measure the relationship between the yield of rice on one hand and the amount of rainfall and temperature taken together, on the other hand. In this case, yield of rice is a dependent variable and rainfall and temperature are independent variables.

Hence, multiple correlation helps us to measure the correlation between a dependent variable and a group of independent variables. The multiple correlation between a dependent variable X_1 and the group of independent variables, X_2 and X_3, is denoted by $R_{1.23}$. The variable before the dot represents independent variable and the variables after the dot represent the independent variables.

4.3.2. Properties of Multiple Correlation

Following are the properties of multiple correlation:

1) Its value lies between 0 and 1 and it never results in a negative figure. Thus, $0 \le R \le 1$.
2) The position of the subscripts to the right of the dot does not make any difference. Thus, $R_{1.23} = R_{1.32}$, and $R_{2.134} = R_{2.341} = R_{2.431}$
3) If $r_{12} = 0$, and $r_{13} = 0$, then $R_{1.23} = 0$ and *vice versa*.
4) $R_{1.23} \ge r_{12}$, and r_{13}; $R_{2.13} \ge r_{12}$ and r_{23}, and $R_{3.12} \ge r_{13}$ and r_{23}. This property is helpful for testing the consistency of the data.
5) Coefficient of a multiple determination is calculated by the square of the coefficient of a multiple correlation

Thus, $R^2 = \frac{\text{Explained variation}}{\text{Total variation}}$

= Coefficient of Multiple Determination

4.3.3. Methods of Calculation

The method for calculating multiple correlation coefficient is based on the number of independent variables. If the number of variables is more than three, then it involves a lot of calculation and hence may require a computer. If the number of variables is less than three, then the multiple correlation can be calculated manually, by using appropriate models, as given below:

1) **When the Total Number of Independent Variables is Two:** In this case, we have three multiple correlation coefficients, as:
   i) $R_{1.23}$, ii) $R_{2.13}$, and iii) $R_{3.12}$

   These will be calculated with the use of following formulae:
   i) Multiple correlation between dependent variable X_1 and the group of independent variables X_2 and X_3 is calculated as follows:

$$R_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - r_{12}\cdot r_{13}\cdot r_{23}}{1 - r_{23}^2}} \text{ or } = \sqrt{r_{12}^2 + r_{13.2}^2\left(1 - r_{12}^2\right)}$$

   ii) Multiple correlation between dependent variable X_2 and the group of independent variables X_1 and X_3 is calculated as follows:

$$R_{2.13} = \sqrt{\frac{r_{12}^2 + r_{23}^2 - 2r_{12}\cdot r_{13}\cdot r_{23}}{1 - r_{13}^2}} \text{ or } = \sqrt{r_{12}^2 + r_{23.1}^2\left(1 - r_{12}^2\right)}$$

   iii) Multiple correlation between dependent variable X_3 and the group of independent variables X_1 and X_2 is calculated as follows:

$$R_{3.12} = \sqrt{\frac{r_{13}^2 + r_{23}^2 - 2r_{12}\cdot r_{13}\cdot r_{23}}{1 - r_{12}^2}} \text{ or } = \sqrt{r_{13}^2 + r_{23.1}^2\left(1 - r_{13}^2\right)}$$

2) **When the Total Number of Independent Variables is Three:** In this case, we have four multiple correlation coefficients, as:
   i) $R_{1.234}$, ii) $R_{2.134}$, iii) $R_{3.124}$, and iv) $R_{4.123}$.

Scanned with CamScanner

These can be calculated with the use of following formulae:

i) $R_{1.234}=\sqrt{1-(1-r_{14}^2)(1-r_{13.4}^2)(1-r_{12.34}^2)}$

ii) $R_{2.134}=\sqrt{1-(1-r_{24}^2)(1-r_{23.4}^2)(1-r_{12.34}^2)}$

iii) $R_{3.124}=\sqrt{1-(1-r_{34}^2)(1-r_{23.4}^2)(1-r_{13.24}^2)}$, and

iv) $R_{4.123}=\sqrt{1-(1-r_{34}^2)(1-r_{24.3}^2)(1-r_{14.23}^2)}$

4.3.4. Advantages of Multiple Correlation

Following are the advantages of multiple correlation analysis:

1) It provides all the utilities which can be availed from partial correlation analysis.
2) It helps us to know the combined effect of all the independent variables on a dependent variable, at the same time.
3) It helps us to test the consistency of the data relating to simple correlation among the various variables.
4) It provides a tool to understand the movements of a dependent variable occurring due to independent variables.
5) It is used as a measure of goodness of fit.
6) It serves as a measure of goodness of fit of regression line and its reliability can be tested at a desired degree of confidence.
7) When studied in conjunction with the partial correlation analysis, it makes a very useful analysis of the relationship between the different variables.

4.3.5. Disadvantages of Multiple Correlation

Following are the disadvantages of multiple correlation analysis:

1) Multiple correlation analysis is based on the assumption that all independent variables are linearly related with dependent variable. Thus, it measures only linear relationship between dependent variable and a group of independent variables. But in practice, most relationships do not follow this assumption; rather they follow some other pattern.
2) It is also assumed that the effect of independent variables on the dependent variable is separate, distinct and additive. When the effects of variables are additive, a given change in one has the same effect on the dependent variable, regardless of the sizes of the other two independent variables.
3) There is a possibility of misinterpretation of the results in case of misuse of correlation.

Example 8: The following correlation coefficients are given for a tri-variate data:

$r_{12}=0.96$, $r_{13}=0.49$ and $r_{23}=0.46$.

Calculate the multiple correlation coefficients treating the first variable as dependent and the second and third variables as independent.

Solution: $R_{1.23}$ = Multiple correlation between X_1 on the one hand, and X_2 and X_3 on the other is given by equation

$$R_{1.23}=\sqrt{\frac{r_{12}^2+r_{13}^2-2r_{12}r_{13}r_{23}}{1-r_{23}^2}}=\sqrt{\frac{(0.96)^2+(0.49)^2-2(0.96)(0.49)(0.46)}{1-(0.46)^2}}$$

$$=\sqrt{\frac{0.9216+0.2401-0.432768}{0.7884}}=\sqrt{\frac{0.728932}{0.7884}}=0.96$$

Example 9:

1) Given that $r_{12}=0.62$, $r_{13}=0.69$ and $r_{23}=0.59$ determine $R_{1.23}$.
2) If $R_{1.23}=0$ does it follow that $R_{2.13}=0$.
3) If $R_{1.23}=1$ does it follow that $R_{2.13}=1$.

Solution:

1) $$R_{1.23}=\sqrt{\frac{r_{12}^2+r_{13}^2-2r_{12}r_{13}r_{23}}{1-r_{23}^2}}=\sqrt{\frac{(0.62)^2+(0.69)^2-2(0.62)(0.71)(0.59)}{1-(0.59)^2}}$$

$$=\sqrt{\frac{0.3844+0.4761-0.519436}{0.6519}}=\sqrt{\frac{0.341064}{0.6519}}=0.723$$

$$r_{23.1}=\frac{r_{23}-r_{12}r_{13}}{\sqrt{(1-r_{12}^2)(1-r_{13}^2)}}=\frac{0.59-(0.62)\times(0.69)}{\sqrt{\{1-(0.62)^2\}\{1-(0.69)^2\}}}=\frac{0.1622}{0.5679}=0.286$$

2) If $R_{1.23}=0$ then $r_{12}^2+r_{13}^2-2r_{12}r_{13}r_{23}=0$ i.e., $r_{12}^2+r_{13}^2-2r_{12}r_{13}r_{23}=0$

The value of $R_{2.13}^2=\frac{r_{12}^2+r_{23}^2-2r_{12}r_{13}r_{23}}{1-r_{13}^2}$

Putting in above, we get: $R_{2.13}^2=\frac{r_{12}^2+r_{23}^2-r_{12}^2-r_{13}^2}{1-r_{13}^2}=\frac{r_{23}^2-r_{13}^2}{1-r_{13}^2}$

So it is not necessary that this value is zero. In fact it will have some positive value.

3) If $R_{1.23}^2=1$ then $r_{12}^2+r_{13}^2-2r_{12}r_{13}r_{23}=1-r_{23}^2$

or $r_{12}^2+r_{23}^2=1-r_{13}^2+2r_{12}r_{13}r_{23}$

The value of $R_{2.13}^2=\frac{1-r_{13}^2+2r_{12}r_{13}r_{23}-2r_{12}r_{13}r_{23}}{1-r_{13}^2}$

Hence if $R_{1.23}=1$, then $r_{2.13}$ will definitely be 1.

4.4. PHARMACEUTICALS EXAMPLES

Example 10: Calculate Karl-Pearson's coefficient of correlation.

X	20	22	24	25	25	26	27	29	30	32
Y	38	40	37	35	33	34	35	36	27	25

Solution:

Calculation of Coefficient of Correlation

X	Y	$x=(X-\bar{X})$	$y=(Y-\bar{Y})$	x^2	y^2	xy
20	38	-6	4	36	16	-24
22	40	-4	6	16	36	-24
24	37	-2	3	4	9	-6
25	35	-1	1	1	1	-1
25	33	-1	-1	1	1	1
26	34	0	0	0	0	0
27	35	1	1	1	1	1
29	36	3	2	9	4	6
30	27	4	-7	16	49	-28
32	25	6	-9	36	81	-54
$\Sigma x=260$	$\Sigma y=340$	0		120	198	-129

Scanned with CamScanner

$$\bar{X} = \frac{\Sigma X}{N} = \frac{260}{10} = 26$$

$$\bar{Y} = \frac{\Sigma Y}{N} = \frac{340}{10} = 34$$

Karl Pearson, coefficient of correlation

$$(r) = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \times \Sigma y^2}} = \frac{-129}{\sqrt{120 \times 198}} = \frac{-129}{\sqrt{23760}} = \frac{-129}{154.14} = -0.836$$

Example 11: Total supply and price per unit of a product are given below. Calculate the Karl Pearson's correlation coefficient:

Supply (in units)	150	400	700	1100	850	900	300
Price per unit	12	23	38	63	45	48	18

Solution: Let assume mean for X is 800 and for Y is 50.

Calculation of Coefficient of Correlation

X	Y	$d_x = (X - A_x)$	$d_y = (Y - A_y)$	d_x^2	d_y^2	$d_x d_y$
150	12	– 650	– 38	422500	1444	24700
400	23	– 400	– 27	160000	729	10800
700	38	– 100	– 12	10000	144	1200
1100	63	300	13	90000	169	3900
850	45	50	– 5	2500	25	– 250
900	48	100	– 2	10000	4	– 200
300	18	– 500	– 32	250000	1024	16000
Total		$\Sigma d_x = -1200$	$\Sigma d_y = -103$	$\Sigma d_x^2 = 945000$	$\Sigma d_y^2 = 3539$	$\Sigma d_x d_y = 56150$

Karl Pearson coefficient of correlations (r)

$$= \frac{N\sum d_x d_y - \sum d_x \sum d_y}{\sqrt{\left\{N\sum d_x^2 - (\sum d_x)^2\right\} \times \left\{N\sum d_y^2 - (\sum d_y)^2\right\}}} = \frac{7(56150) - (-1200 \times -103)}{\sqrt{7 \times 945000 - (-1200)^2} \times \sqrt{7 \times 3539 - (-103)^2}}$$

$$= \frac{393050 - 123600}{\sqrt{6615000 - 1440000} \times \sqrt{24773 - 10609}} = \frac{269450}{\sqrt{5175000} \times \sqrt{14164}} = \frac{269450}{270725 \times 119} = 0.995$$

Example 12: A researcher wished to determine if a person's age is related to the number of hours he or she exercises per week. The data obtained from a sample is given. State your opinion based on Karl Pearson's coefficient of correlation for the data.

Age	16	24	38	46	54	62
Hours	11	7	5	4	2	1

Solution:

Calculation of Coefficient of Correlation

Age (X)	Hours (Y)	X^2	Y^2	XY
16	11	256	121	176
24	7	576	49	168
38	5	1444	25	190
46	4	2116	16	184
54	2	2916	4	108
62	1	3844	1	62
$\Sigma X = 240$	$\Sigma Y = 30$	$\Sigma X^2 = 11152$	$\Sigma Y^2 = 216$	$\Sigma XY = 888$

Coefficient of Correlation (r)

$$= \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{\sqrt{N\Sigma X^2 - (\Sigma X)^2}\sqrt{N\Sigma Y^2 - (\Sigma Y)^2}} = \frac{6 \times 888 - (240)(30)}{\sqrt{6 \times 11152 - (240)^2}\sqrt{6 \times 216 - (30)^2}}$$

$$= \frac{5328 - 7200}{\sqrt{66912 - 57600} \times \sqrt{1296 - 900}} = \frac{-1872}{\sqrt{9312} \times \sqrt{396}} = \frac{-1872}{1920.3} = -0.975$$

So there is a **negative correlation** between age and number of hours he/she exercises per week.

Example 13: Marks obtained by ten students in Pharmacology and Pharmaceutical chemistry are shown in table below:

Pharmacology	8	36	98	25	75	82	92	62	65	39
Pharmaceutical Chemistry	84	51	91	60	68	62	86	58	35	49

Determine the rank correlation coefficient.

Solution:

$$\text{Mean } \bar{x} = \frac{\Sigma x_i}{n} = \frac{8+36+98+25+75+82+92+62+65+39}{10} = \frac{582}{10} = 58.2$$

$$\text{Mean } \bar{y} = \frac{\Sigma y_i}{n} = \frac{84+51+91+60+68+62+86+58+35+49}{10} = \frac{644}{10} = 64.4$$

Thus, $\bar{x} = 58.2$ is not an integer, we can take assumed mean A = 58

$\bar{y} = 64.4$ is not an integer, we can take assumed mean B = 64

X	Y	Dx = x – A = x – 58	Dy = y – B = y – 64	dx^2	dy^2	dx.dy
8	84	–50	20	2500	400	–1000
36	51	–22	–13	484	169	286
98	91	40	27	1600	729	1080
25	60	–33	–4	1089	16	132
75	68	17	4	289	16	68
82	62	24	–2	576	4	–48
92	86	34	22	1156	484	748
62	58	4	–6	16	36	–24
65	35	7	–29	49	841	–203
39	49	–19	–15	361	225	285
$\Sigma X = 582$	$\Sigma Y = 644$	$\Sigma dx = 2$	$\Sigma dy = 4$	$\Sigma dx^2 = 8120$	$\Sigma dy^2 = 2920$	$\Sigma dx \cdot dy = 1324$

Correlation Coefficient r:

$$r = \frac{n \cdot \Sigma dx dy - \Sigma dx \cdot \Sigma dy}{\sqrt{n \cdot \Sigma dx^2 - (\Sigma dx)^2} \times \sqrt{n \cdot \Sigma dy^2 - (\Sigma dy)^2}} = \frac{10 \cdot 1324 - 2 \cdot 4}{\sqrt{10 \cdot 8120 - (2)^2} \times \sqrt{10 \cdot 2920 - (4)^2}}$$

$$= \frac{13240 - 8}{\sqrt{81200 - 4} \times \sqrt{29200 - 16}} = \frac{13232}{\sqrt{81196} \times \sqrt{29184}} = \frac{13232}{284.9491 \times 170.8333}$$

$$= \frac{13232}{48678.7845} = 0.2718$$

Scanned with CamScanner

Example 14: The Apgar score and birth weight (kg) of 10 infants are shown in table below. Find the correlation coefficient and interpret.

Apgar Score	3	2	3	4	5	2	3	5	4	5
Birth Weight	3.6	2.9	3.2	3.8	3.9	3	3.5	3.8	3.9	3.9

Solution:

x	y	x^2	y^2	x·y
3	3.6	9	12.96	10.8
2	2.9	4	8.41	5.8
3	3.2	9	10.24	9.6
4	3.8	16	14.44	15.2
5	3.9	25	15.21	19.5
2	3	4	9	6
3	3.5	9	12.25	10.5
5	3.8	25	14.44	19
4	3.9	16	15.21	15.6
5	3.9	25	15.21	19.5
$\sum x = 36$	$\sum y = 35.5$	$\sum x^2 = 142$	$\sum y^2 = 127.37$	$\sum xy = 131.5$

Correlation Coefficient r:

$$r = \frac{n \cdot \sum xy - \sum x \cdot \sum y}{\sqrt{n \cdot \sum x^2 - (\sum x)^2} \times \sqrt{n \cdot \sum y^2 - (\sum y)^2}} = \frac{10 \cdot 131.5 - 36 \cdot 35.5}{\sqrt{10 \cdot 142 - (36)^2} \times \sqrt{10 \cdot 127.5 - (35.5)^2}}$$

$$= \frac{1315 - 1278}{\sqrt{1420 - 1296} \times \sqrt{1273.7 - 1260.25}} = \frac{37}{\sqrt{124} \times \sqrt{13.45}} = \frac{37}{11.1355 \times 3.6674} = \frac{37}{40.8387} = 0.906$$

Example 15: Statistician of a tobacco company wants to know how heavy smoking is related to longevity. From the data of a sample of recently deceased smokers, the number of cigarettes (estimated on a per day for their last five years) is paired with the number of years that they lived.

Subject	Cigarettes	Years Lived
1	25	63
2	35	68
3	10	72
4	40	62
5	85	65
6	75	46
7	60	51
8	45	60
9	50	55

Determine rank coefficient of correlattion.

Solution: Let us consider two arbitrarily variable x and y.

The outcome of Pearson's value r will be exactly similar no matter which variable is taken as x or y.

Step 1) First, obtain $\sum x$, $\sum x^2 (\sum x)^2$, and $(\sum y)^2$

$\sum x = 25 + 35 + 10 + 40 + 85 + 75 + 60 + 45 + 50 = 425$

$\sum x^2 = 25^2 + 35^2\ 10^2 + 40^2 + 85^2 + 75^2 + 60^2 + 45^2 + 50^2 = 24{,}525$

$\sum y = 63 + 68 + 72 + 62 + 65 + 46 + 51 + 60 + 55 = 518$

$\sum y^2 = 63^2 + 68^2 + 72^2 + 62^2 + 65^2 + 46^2 + 51^2 + 60^2 + 55^2 = 33{,}188$

$(\sum x)^2 = (25 + 35 + 10 + 40 + 85 + 75 + 60 + 45 + 50)^2$

$(\sum x)^2 = 425^2 = 180{,}625$

$(\sum y)^2 = (63 + 68 + 72 + 62 + 65 + 46 + 51 + 60 + 55)^2$

$(\sum y)^2 = 542^2 = 293{,}764$

Step 2) Find the sum ($\sum xy$) by multiplying each x score by its paired y score.

$\sum xy = (25 \times 63) + (35 \times 68) + (10 \times 72) + (40 \times 62) + (85 \times 65) + (75 \times 46) + (60 \times 51) + (45 \times 60) + (50 \times 55)$

$\sum xy = 1575 + 2380 + \ldots + 2750$

$\sum xy = 24{,}640$

Step 3) First we find the numerator value of formula of r. Here, N shows the number of pairs of scores. In this example, total of 9 pairs of scores.

$N\sum xy - (\sum x)(\sum y) = (9)(24{,}640) - (425)(518) = 22{,}1760 - 230{,}350 = -8{,}590$

Step 4) After calculating the numerator value, we now find the denominator of r formula. Thus, we have denominator value as follows:

$$\sqrt{\left[N\sum x^2 - (\sum x)^2\right]\left[N\sum y^2 - (\sum y)^2\right]} = \sqrt{(9)(24{,}525) - (425)^2]\,[(9)(33{,}188) - (542)^2]}$$

$$= \sqrt{[220{,}725 - 180{,}625]\,[298{,}692 - 293{,}764]} = \sqrt{[40{,}100]\,[4{,}928]} = 14{,}057.482$$

Step 5) After dividing the result of numerator from **Step 3** by the result of denominator of **Step 4**, we get the following:

$$r = \frac{-8{,}590}{14{,}057.482} = -.6111 = -.61$$

The Pearson r is usually rounded off to two decimal places. Because r = –0.61 it means that there is a strong negative correlation between smoking and longevity. This shows that the higher the number of cigarettes smoked in the past five years, the lower the number of years lived. And if less number of cigarettes smoked then the higher the number of year lived.

Example 16: Table 4.2 shows the mean caloric consumption per day *versus* weight change per month.

Table 4.2: Data of Caloric Consumption *versus* Weight Change

Patient	(X) Mean Caloric Consumption/Day	(Y) Weight Change/Month
1	1,200	0.0
2	1,500	0.5
3	1,800	0.5
4	2,000	1.5
5	2,500	4.0
6	1,800	1.0
7	2,500	3.0
8	2,000	2.0

Calculate the Karl Pearson's coefficient of correlation.

Solution: Table 4.3: Calculation of Pearson Product-Moment Correlation Coefficient

Patient	X	$\overline{X}$	$(X_i - \overline{X})$	y	$\overline{Y}$	$(y_i - \overline{Y})$	$(x_i - \overline{X})(y_i - \overline{Y})$
1	1,200	– 1,912.5	–712.5	0.0	1.56	–1.56	1,111.5
2	1,500	– 1,912.5	–412.5	0.5	1.56	–1.06	437.25
3	1,800	– 1,912.5	–112.5	0.5	1.56	–1.06	119.25
4	2,000	– 1,912.5	87.5	1.5	1.56	–0.06	–5.25
5	2,500	– 1,912.5	587.5	4.0	1.56	–2.44	1,433.5
6	1,800	– 1,912.5	–112.5	1.0	1.56	–0.56	–63.0
7	2,500	– 1,912.5	587.5	3.0	1.56	–1.44	846.0
8	2,000	– 1,912.5	87.5	2.0	1.56	–0.44	38.5
	X = 1,912.5			Y = 1.56			$\sum$ 4,043.75

Scanned with CamScanner

$$S_{xy} = \sum[(x_i - \bar{X})(y_i - \bar{Y})]/(n-1) = \frac{4,043.75}{8-1} = 577.7$$

$$S_x = \frac{\sum(x_i - X)^2}{(n-1)} = \frac{1,408,750}{8-1} = 448.6;\ S_y = \frac{\sum(y_i - Y)^2}{(n-1)} = \frac{13.22}{8-1} = 1.37$$

$$r = \frac{S_{xy}}{S_x S_y} = \frac{577.7}{(448.6)(1.37)} = 0.94$$

Karl Pearson Coefficient of r is 0.94. This illustrates some degree of positive correlation.

Example 17: From the data of dry bark (x_1), height (x_2) and girth (x_3) for 18 cinchona plants, the following coefficients of correlation were obtained:
$r_{12} = 0.77$, $r_{13} = 0.72$ and $r_{23} = 0.52$.

Calculate the coefficient of multiple correlation $R_{1.23}$.

Solution: Coefficient of Multiple correlation can be calculated using the following formula:

$$R_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{23}^2}} = \sqrt{\frac{(0.77)^2 + (0.72)^2 - 2\times0.77\times0.72\times0.52}{1-(0.52)^2}} = \sqrt{0.7334} = +0.8564$$

This shows that the multiple correlation coefficient is non-negative.

Example 18: Determine the multiple correlation coefficients $R_{1.23}$. The simple correlation coefficients between temperature (x_1), yield of corn (x_2), and rainfall (x_3) are given as follows:
$r_{12} = 0.59$, $r_{13} = 0.46$, and $r_{23} = 0.77$.

Solution: Multiple correlation coefficient can be calculated as follows:

$$R_{1.32} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{23}^2}} = \sqrt{\frac{(0.59)^2 + (0.46)^2 - 2(0.59\times0.46\times0.77)}{1-(0.77)^2}}$$

$$= \sqrt{\frac{0.3481 + 0.2116 - 0.418}{0.4071}} = \sqrt{\frac{0.5597 - 0.418}{0.4071}} = \sqrt{\frac{0.1417}{0.4071}} = 0.589$$

Example 19: The sample data of fasting blood glucose level and systolic blood pressure in 10 diabetic patients are shown in table below:

Sr. no of patients	1	2	3	4	5	6	7	8	9	10
Fasting Blood Glucose Level (mg/dl)	90	92	98	112	120	121	126	132	143	145
Systolic Blood Pressure (mmHg)	136	140	142	130	148	135	150	170	145	165

Solution:

$$\text{Mean } \bar{x} = \frac{\sum x_i}{n} = \frac{90+92+98+112+120+121+126+132+143+145}{10} = \frac{1179}{10} = 117.9$$

$$\text{Mean } \bar{y} = \frac{\sum y_i}{n} = \frac{136+140+142+130+148+135+150+170+145+165}{10} = \frac{1461}{10} = 146.1$$

$\bar{x} = 117.9$ is not an integer, we use assumed mean A = 118
$\bar{y} = 146.1$ is not an integer, we use assumed mean B = 146.

x	y	$dx = x - A = x - 118$	$dy = y - B = y - 146$	dx^2	dy^2	$dx \cdot dy$
90	136	−28	−10	784	100	280
92	140	−26	−6	676	36	156
98	142	−20	−4	400	16	80
112	130	−6	−16	36	256	96
120	148	2	2	4	4	4
121	135	3	−11	9	121	−33
126	150	8	4	64	16	32
132	170	14	24	196	576	336
143	145	25	−1	625	1	−25
145	165	27	19	729	361	513
–	–	–	–	–	–	–
1179	1461	$\sum dx = -1$	$\sum dy = 1$	$\sum dx^2 = 3523$	$\sum dy^2 = 1487$	$\sum dx \cdot dy = 1439$

Correlation Coefficient r:

$$r = \frac{n\cdot\sum dxdy - \sum dx\cdot\sum dy}{\sqrt{n\cdot\sum dx^2 - (\sum dx)^2}\cdot\sqrt{n\cdot\sum dy^2 - (\sum dy)^2}}$$

$$= \frac{10\cdot1439 - (1\cdot1)}{\sqrt{10\cdot3523 - (-1)^2}\cdot\sqrt{10\cdot1487 - (1)^2}} = \frac{14390+1}{\sqrt{35230-1}\cdot\sqrt{14870-1}} = \frac{14391}{\sqrt{35229}\cdot\sqrt{14869}}$$

$$= \frac{14391}{187.6939\cdot121.9385} = \frac{14391}{22887.1143} = 0.6288$$

Considering appendix for n=1, we calculate $0.02 < P < 0.05$. So, null hypothesis of no correlation is rejected and it is determined that fasting blood glucose and systolic blood pressure remain correlated.

4.5. EXERCISE

4.5.1. Very Short Answer Type Questions

1) What is correlation?
2) What is linear correlation?
3) Define scatter or dot diagram.
4) Write the properties of Karl Pearson's coefficient of correlation.
5) Define the multiple correlation and example.

4.5.2. Short Answer Type Questions

1) Explain the various type of correlation.
2) What is degree of correlation and coefficient correlation?
3) Discuss about scatter or dot diagram with example.
4) Calculate Range and its Coefficient from the following data:
   53, 46, 18, 16, 75, 84, 28
   [Ans: Coefficient of range = 0.65]
5) Write the advantages and disadvantages for multiple correlations.
6) The mean of two samples of size 50 and 100 respectively is 54.1 and 50.3 and the standard deviations are 8 and 7. Obtain the standard deviation of the sample of size 150 obtained by combining the two samples.
   [Ans: Combined σ= 7.5636]
7) The standard deviation of a distribution of 100 values is ₹2. If the sum of the squares of the actual values is ₹3,600, what is the mean of this distribution?
   [Ans: 5.66]

Scanned with CamScanner

4.5.3. Long Answer Type Questions

1) Find Karl Pearson's coefficients of correlation from the following index numbers and interpret it:

Wages	100	101	103	102	104	99	97	98	96	96
Cost of Living	98	99	99	97	95	02	95	94	90	91

[Ans: r = 0.85]

2) Calculate the Karl Pearson's Coefficient of Correlation between age and playing habits from the data given below. Comment on the value:

Age	20	21	22	23	24	25
No. of Students	500	400	300	240	200	160
Regular Players	400	300	180	96	60	24

[Ans: r = – 0.991]

3) Find Rank correlation coefficient:

Marks in Economics	80	45	55	56	58	60	65	68	20	75	85
Marks in Mathematics	82	86	50	48	60	62	64	65	70	74	90

[Ans: R = 0.58]

4) Calculate coefficient of correlation between marks in the subject of maths and statistics.

Maths Marks	:	40,	34,	28,	30,	44,	38	31	
Statistics Marks	:	32,	39,	26,	30,	38,	34,	28	

[Ans: r = 0.705]

5) In a fancy dress competition, two judges accorded following ranks to the 10 participants:

Judge X	2	9	7	5	3	10	1	6	8	4
Judge Y	10	8	1	4	6	9	7	2	5	3

Calculate Coefficient of Rank Correlation.
[Ans: R = 0.54]

6) Two departmental managers ranked a few trainees according to their perceived abilities. The ranks are given as:

Trainees	A	B	C	D	E	F	G	H	I	K
Manager A	1	9	6	2	5	8	7	3	10	4
Manager B	3	10	8	1	7	5	6	2	9	4

Calculate an appropriate correlation coefficient to measure consistency in ranking.
[Ans: R = 0.84]

CHAPTER 5 Regression

5.1. REGRESSION

5.1.1. Meaning and Definition

Regression analysis is a statistical tool used to calculate a continuous dependent variable from various independent variables and is commonly used for prediction and forecasting. **For example,** if agriculturist knows that the yield of rice and rainfall are closely related then he will want to know the amount of rain required to achieve a certain production. In this situation, he will use the regression analysis.

This technique is used for modelling and analysis of mathematical data using the value of a **dependent variable** and one or more **independent variables.** Dependent variable, also known as **response variable,** is the single variable predicted by the regression model. Independent variable, also known as **predictor variable,** is used to predict the values of dependent variable.

Regression analysis is the process to calculate the unknown value of one variable from the known values of the other variable.

According to Blair, "Regression is the measure of the average relationship between two or more variable in terms of the original units of the data."

According to Taro Yamane, "One of the most frequently used techniques in economics and business research, to find a relation between two or more variables that are related causally, is regression analysis."

Regression analysis makes an equation to describe the statistical relationship between one or more predictor variables and the response variable. It helps to know, how to determine the nature and strength of a relatonship between independent and dependent variables.

Regression may be classified as follows:
1) Simple Regression
2) Multiple Regreesion
3) Linear Regreesion
4) Non-linear Regression

5.1.2. Application of Regression Analysis

Following are the basic uses of regression analysis:
1) Used to know the relationship between different (one or more) variables.
2) Used to find out the value of dependent variable from the value of independent variables.
3) Used to find out the coofficients of correlation (r) and coofficients of determination.
4) In corporate sector it is useful to check the quality.
5) Also very useful to determine the statistical curve (demand, supply, etc.).

Scanned with CamScanner

5.1.3. Difference between Correlation and Regression Analysis

Table 5.1 shows the difference between Correlation and Regression Analysis:

Table 5.1: Difference between Correlation and Regression Analysis

Basis	Correlation	Regression
Meaning	Study of the linear relationship between two variables is known as correlation.	It is a statistical tool used to calculate a continuous dependent variable from various independent variables.
Variables	Random variables are used.	One variable is random variable and other is fixed variable. It may be possible that sometime both the variables are random.
Relationship	It does not indicate the cause and effect relationship between variables and gives the degree of relationship.	It shows the cause and effect relationship between variables and creates a functional relationship.
Uses	Used to test the relationship between two variables.	Used for prediction and forecasting.
Value/ Range	It is a relative measure and its coefficient is lies between –1 to +1.	It is an absolute value. It is possible that the coefficient is greater than 1.
Mathematical Treatment	Not much valuable for further mathematical treatment.	Widely used for further mathematical treatment.
Dependency of Variables	Dependency of variables on each other is not important.	Its shows the functional relationship between the two variables.

5.1.4. Regression Lines

The lines of the best fit between two variables and stating the common average relationship are known as regression lines. There are two regression lines for two variables. When we take two variables X as dependent variable and Y as independent variable, then the straight line is called as the **regression line of X and Y** and *vice versa* is also true.

The regression line of **Y on X** gives the finest possible values of Y for the given values of X. So this line is the sum of squares of deviations of the 'calculated values of Y' and 'observed values of Y is minimum'; i.e.,

$$\Sigma(Y_c - Y_0)^2 = \text{minimum}$$

Where, Y_c = Calculated value of Y
Y_0 = Observed value of Y

Now, for the regression line of **X on Y**:

$$\Sigma(X_c - X_0)^2 = \text{minimum.}$$

Where, X_c = Calculated value of X
X_0 = Observed value of X

This is why, to understand the concept of regression line, two regression lines are required. If there is only one regression line existing between two variables then it means $r = \pm 1$.

Now If the value of r is 0, then both the variance are independent and lines will cut each other at a right angle (i.e., lines will be parallel (ll) to x-axis and y-axis) (**figure 8.1**).

5.1.5. Regression Equations

Regression equation is used to show the regression line and there are two regression equations because there are two regression lines.

So if there are two variables, say X and Y, then the two regression equations are written as:

1) **Regression Equation of X on Y:** The regression equation of 'X on Y' is written as follows:

$$X = a + bY$$

To calculate the values of a and b, following normal equations are solved simultaneously:

$$\Sigma X = Na + b\Sigma Y; \quad \Sigma XY = a\Sigma Y + b\Sigma Y^2$$

2) **Regression Equation of Y on X:** The regression equation of 'Y on X' is written as follows:

$$Y = a + bX$$

To get the value of a and b, following normal equations are to be solved simultaneously:

$$\Sigma Y = Na + b\Sigma X; \quad \Sigma XY = a\Sigma X + b\Sigma X^2$$

In the normal equation ΣX, ΣXY, ΣX^2 and ΣY^2 gives the summations and can be determined from the observed pairs of values of two variables 'X and Y'. N is the number of observed pairs of values.

5.1.6. Regression Coefficient

If 'b' is the slope of regression line of 'Y on X', then b_{YX} is known as coefficient of regression of 'Y on X'.

$$\text{So, } b_{YX} = \text{Regression Coefficient of Y on X} = r\frac{\sigma_Y}{\sigma_X}$$

Similarly, for the regression line of 'X on Y',

$$b_{XY} = \text{Regression Coefficient of X on Y} = r\frac{\sigma_X}{\sigma_Y}$$

Properties of Regression Coefficients

1) Coefficient of correlation is calculated by determining the geometric mean of the two regression coefficients i.e.,

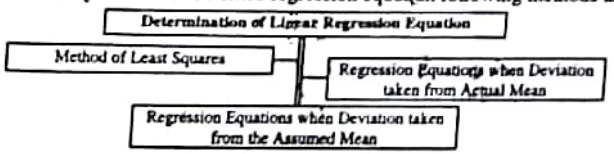
$$r = \pm\sqrt{b_{XY} \times b_{YX}}$$

Where r = Coefficient of correlation and its value is positive when both the regression coefficients are positive. Because r lies between +1 to –1, so the value of regression coefficient is not more than one.

2) It is independent from origin but not from scale.
3) If one regression coefficient is greater than unity, then remaining must be smaller than unity.
4) Both the regression coefficients have the same sign (i.e., if one is positive then the other is also positive and if one is negative then the other is also negative).
5) The value of coefficient of correlation is less than the mean value of regression coefficients.

5.1.7. Determination of Linear Regression Equation

To determine the parameters of a fitted regression equation following methods are used.



Scanned with CamScanner

5.1.7.1. Curve Fitting by the Method of Least Squares (Fitting the Lines y = a + bx and x = a + by)

1) Let X = a + bY be the line of X on Y. To determine the values of a and b, the following two normal equations are to be solved simultaneously:

$\Sigma X = Na + b\Sigma Y$

$\Sigma XY = a\Sigma Y + b\Sigma Y^2$

Example 1: Consider the following data obtain the two regression equations:

X	5	3	10	8	9
Y	8	12	6	5	14

Solution:

Calculation of Regression Equations

X	Y	X^2	Y^2	XY
5	8	25	64	40
3	12	9	144	36
10	6	100	36	60
8	5	64	25	40
9	14	81	196	126
$\Sigma X = 35$	$\Sigma Y = 45$	$\Sigma X^2 = 279$	$\Sigma Y^2 = 465$	$\Sigma XY = 302$

Regression equation of **X on Y:** X = a + bY and the two normal equations are:

$\Sigma X = Na + b\Sigma Y$

$\Sigma XY = a\Sigma Y + b\Sigma Y^2$

Substituting the values

$35 = 5a + 45b$ ..... (1)

$302 = 45a + 465b$ ..... (2)

Multiplying equation (1) by 9:

$315 = 45a + 405b$ ..... (3)

Subtract equation (2) from (3),

$13 = -60b$ or $b = -0.22$

Substituting the value of b in equation (1):

$35 = 5a + 45(-0.22)$

$\Rightarrow 35 = 5a - 9.9$

$\Rightarrow 5a = 35 + 9.9 = 44.9$

$\therefore a = 8.98$

Putting the value of a and b in the equation,
The regression equation of **X on Y:**

$X = 8.98 - 0.22\,Y$.

2) Let Y = a + bX be the line of Y on X. To determine the values of a and b, the following two normal equations are to be solved simultaneously

$\Sigma Y = Na + b\Sigma X$

$\Sigma XY = a\Sigma X + b\Sigma X^2$

Substituting the values,

$45 = 5a + 35b$ ..... (1)

$302 = 35a + 279b$ ..... (2)

Multiplying equation (1) by 7,

$315 = 35a + 245b$ ..... (3)

Subtracting equation (2) from (3),

$13 = -34b$ or $b = -0.38$

Substituting the value of b in equation (1),

$45 = 5a + 35(-0.38)$

$5a = 45 + 13.3 = 58.3$ or $a = 11.66$

Putting the values of a and b in the equation
Regression equation of Y on X:

$Y = 11.66 - 0.38\,X$

5.1.7.2. Regression Equations when Deviation Taken from Actual Mean

Following formulae are used to find out the deviation of X and Y series when deviation is taken from the actual mean.

1) **Regression Equation of X on Y**

$$X - \bar{X} = r\frac{\sigma_x}{\sigma_y}(Y - \bar{Y})$$

$\bar{X}$ = Mean of X Series and $\bar{Y}$ = Mean of Y Series

$r\frac{\sigma_X}{\sigma_Y}$ = Regression Coefficient of X on Y or b_{XY}

$$b_{XY} = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{\sum(Y - \bar{Y})^2} = \frac{\sum xy}{\sum y^2}$$

Or $$\frac{\sum XY - \frac{(\sum X)(\sum Y)}{N}}{\sum Y^2 - \frac{(\sum Y)^2}{N}}$$

Where, $x = (X - \bar{X})$; $y = (Y - \bar{Y})$

2) **Regression Equation of Y on X**

$$Y - \bar{Y} = r\frac{\sigma_Y}{\sigma_X}(X - \bar{X})$$

$\bar{X}$ = Mean of X Series and $\bar{Y}$ = Mean of Y Series

$r\frac{\sigma_Y}{\sigma_X}$ = The Regression Coefficient of Y on X = b_{YX}.

$$b_{YX} = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{\sum(X - \bar{X})^2} = \frac{\sum xy}{\sum x^2} \text{ Or } \frac{\sum XY - \frac{(\sum X)(\sum Y)}{N}}{\sum X^2 - \frac{(\sum X)^2}{N}}$$

Where, $x = (X - \bar{X})$; $y = (Y - \bar{Y})$

Example 2: Using the following data obtain two regression equations.

X	15	19	24	21	25	22	16	20	27	21
Y	31	36	48	37	50	45	33	41	39	20

Scanned with CamScanner

Solution: **Calculation of Regression Equations**

X	Y	$x = X - \bar{X}$	$y = Y - \bar{Y}$	x^2	y^2	xy
15	31	-6	-7	36	49	42
19	36	-2	-2	4	4	4
24	48	3	10	9	100	30
21	37	0	-1	0	1	0
25	50	4	12	16	144	48
22	45	1	7	1	49	7
16	33	-5	-5	25	25	25
20	41	-1	3	1	9	-3
27	39	6	1	36	1	6
21	20	0	-18	0	324	0
$\Sigma X = 210$	$\Sigma Y = 380$	$\Sigma x = 0$	$\Sigma y = 0$	$\Sigma x^2 = 128$	$\Sigma y^2 = 706$	$\Sigma xy = 159$

$$\bar{X} = \frac{210}{10} = 21 \quad \text{and} \quad \bar{Y} = \frac{380}{10} = 38$$

$$b_{XY} = \frac{\sum xy}{\sum y^2} = \frac{159}{706} = 0.225; \qquad b_{YX} = \frac{\sum xy}{\sum x^2} = \frac{159}{128} = 1.242$$

Regression equation of X on Y is:

$$X - \bar{X} = b_{XY}(Y - \bar{Y})$$

$\Rightarrow X - 21 = 0.225(Y - 38)$
$\Rightarrow X = 0.225Y - 38 \times 0.225 + 21$
$\Rightarrow X = 0.225Y - 8.55 + 21$
$\Rightarrow X = 12.45 + 0.225Y$

Regression equation of Y on X is:

$$Y - \bar{Y} = b_{YX}(X - \bar{X})$$

$\Rightarrow Y - 38 = 1.242(X - 21)$
$\Rightarrow Y = 1.242X - 21 \times 1.242 + 38$
$\Rightarrow Y = 1.242X - 26.1 + 38$
$\Rightarrow Y = 11.9 + 1.242X$

5.1.7.3. Regression Equations when Deviation Taken from the Assumed Mean

If we take the deviation from the assumed mean then the following formulae are used to solve the regression equation.

Note: Assumed mean is taken because value of actual mean is in the fraction form.

Regression Equation of X on Y is:

$$X - \bar{X} = r\frac{\sigma_X}{\sigma_Y}(Y - \bar{Y})$$

Now find out the value of $r\frac{\sigma_X}{\sigma_Y}$ by applying the following formula:

$$b_{XY} \text{ or } r\frac{\sigma_X}{\sigma_Y} = \frac{\Sigma d_x d_y - \frac{\Sigma d_x \times \Sigma d_y}{N}}{\Sigma d_y^2 - \frac{(\Sigma d_y)^2}{N}} \text{ or } \frac{N\Sigma d_x d_y - \Sigma d_x \Sigma d_y}{N\Sigma d_y^2 - (\Sigma d_y)^2}$$

Regression Equations of Y on X is:

$$Y - \bar{Y} = r\frac{\sigma_Y}{\sigma_X}(X - \bar{X})$$

Now find out the value of $r\frac{\sigma_Y}{\sigma_X}$ by applying the following formula:

$$b_{YX} \text{ or } r\frac{\sigma_Y}{\sigma_X} = \frac{\Sigma d_x d_y - \frac{\Sigma d_x \times \Sigma d_y}{N}}{\Sigma d_x^2 - \frac{(\Sigma d_x)^2}{N}} \text{ or } \frac{N\Sigma d_x d_y - \Sigma d_x \Sigma d_y}{N\Sigma d_x^2 - (\Sigma d_x)^2}$$

Where, $d_x = X - A_x$, $d_y = Y - A_y$

Example 3: From the following data find the regression equations, and estimate the likely value of Y when X = 100.

X	74	98	76	82	58	76	94	88	51
Y	124	131	117	131	96	121	136	97	85

Solution: Let assume that the mean for X-series is 80 and for Y-series is 120.

X	Y	$d_x = (X - A_x)$	$d_y = (Y - A_y)$	$d_x d_y$	d_x^2	d_y^2
74	124	-6	4	-24	36	16
98	131	18	11	198	324	121
76	117	-4	-3	12	16	9
82	131	2	11	22	4	121
58	96	-22	-24	528	484	576
76	121	-4	1	-4	16	1
94	136	14	16	224	196	256
88	97	8	-23	-184	64	529
51	85	-29	-35	1015	841	1225
$\Sigma X = 697$	$\Sigma Y = 1038$	$\Sigma d_x = -23$	$\Sigma d_y = -42$	$\Sigma d_x d_y = 1787$	$\Sigma d_x^2 = 1981$	$\Sigma d_y^2 = 2854$

$$\bar{X} = \frac{697}{9} = 77.44 \quad \text{and} \quad \bar{Y} = \frac{1038}{9} = 115.33$$

$$b_{XY} = \frac{N\sum d_x d_y - \sum d_x \sum d_y}{N\sum d_y^2 - (\sum d_y)^2} = \frac{9 \times 1787 - (-23)(-42)}{9 \times 2854 - (-42)^2}$$

$$= \frac{16083 - 966}{25686 - 1764} = \frac{15117}{23922} = 0.632$$

$$b_{YX} = \frac{N\sum d_x d_y - \sum d_x \sum d_y}{N\sum d_x^2 - (\sum d_x)^2} = \frac{9 \times 1787 - (-23)(-42)}{9 \times 1981 - (-23)^2} = \frac{16083 - 966}{17829 - 529} = \frac{15117}{17300} = 0.874$$

Regression equation of X on Y is:

$$X - \bar{X} = b_{XY}(Y - \bar{Y})$$

$\Rightarrow X - 77.44 = 0.632(Y - 115.33)$
$\Rightarrow X - 77.44 = 0.632Y - 115.33 \times 0.632$
$\Rightarrow X - 77.44 = 0.632Y - 72.88$
$\Rightarrow X = 0.632Y - 72.88 + 77.44$
$\Rightarrow X = 4.56 + 0.632Y$

Regression equation of Y on X is:

$$Y - \bar{Y} = b_{YX}(X - \bar{X})$$

$\Rightarrow Y - 115.33 = 0.874\,(X - 77.44)$
$\Rightarrow Y - 115.33 = 0.874X - 77.44 \times 0.874$
$\Rightarrow Y - 115.33 = 0.874X - 67.68$
$\Rightarrow Y = 0.874X - 67.68 + 115.33$
$\Rightarrow Y = 47.65 + 0.874\,X$

When X = 100, Y = 47.65 + 0.874 × 100 = 47.65 + 87.4 = 135.05

Scanned with CamScanner

5.1.8. Relationship between Correlation & Regression Coefficients

The correlation coefficient is the **Geometric mean** between the regression coefficients, i.e.,

$$r^2 = b_{yx} \times b_{xy}$$

$$\Rightarrow \quad r = \sqrt{b_{yx} \times b_{xy}} \qquad \text{.....(1)}$$

Let us suppose that the two lines of regression are:

$$a_1X + b_1Y + c_1 = 0 \qquad \text{.....(2)}$$

$$a_2X + b_2Y + c_2 = 0 \qquad \text{.....(3)}$$

Let us suppose that equation (2) is the line of regression of Y on X and equation (3) is the line of regression of X on Y.

From equation (2),

$$a_1X + b_1Y + c_1 = 0$$

Or $$b_1Y = -a_1X - c_1$$

Or $$Y = \frac{-a_1}{b_1}X - \frac{c_1}{b_1} \Rightarrow b_{yx} = \frac{-a_1}{b_1}$$

From equation (3),

$$a_2X + b_2Y + c_2 = 0 \qquad \text{or} \qquad a_2X = -b_2Y - c_2$$

or $$X = \frac{-b_2}{a_2}Y - \frac{c_2}{a_2} \Rightarrow b_{xy} = \frac{-b_2}{a_2}$$

The correlation coefficient r between X and Y can now be obtained by this formula:

$$r^2 = b_{yx} \cdot b_{xy} = \left(\frac{-a_1}{b_1}\right) \times \left(\frac{-b_2}{a_2}\right) = \frac{a_1b_2}{b_1a_2}$$

$$\therefore \quad r = \pm\sqrt{\frac{a_1b_2}{a_2b_1}} \qquad \text{.....(4)}$$

Example 4: If the two lines of regression are:

$4X - 5Y + 30 = 0$ and $20X - 9Y - 107 = 0$,

Which of these is the line of regression of X on Y, and Y on X? Find r_{xy} and σ_y when $\sigma_x = 3$.

Solution: We are given the regression lines as:

$$4X - 5Y + 30 = 0 \qquad \text{.....(1)}$$

and $$20X - 9Y - 107 = 0 \qquad \text{.....(2)}$$

Let (1) be the equation of the line of regression of X on Y and (2) be the equation of the line of regression of Y on X.

From (1), we get $X = \frac{5}{4}Y - \frac{30}{4} \Rightarrow b_{XY} = \frac{5}{4}$

From (2), we get $Y = \frac{20}{9}X - \frac{107}{9} \Rightarrow b_{YX} = \frac{20}{9}$

$$\therefore \quad r^2 = b_{YX} \cdot b_{XY} = \frac{20}{9} \times \frac{5}{4} = 2.7778$$

Since $r^2 > 1$, our supposition is wrong.

[$\because$ We always have $0 \le r^2 \le 1$]

Hence, (1) is the line of regression of Y on X and (2) is the line of regression of X on Y.

Re-writing (1), we get, $5Y = 4X + 30 \Rightarrow Y = \frac{4}{5}X + 6$

$$\Rightarrow b_{YX} = \frac{4}{5}$$

Re-writing (2), we get, $20X = 9Y + 107$

$$\Rightarrow X = \frac{9}{20}Y + \frac{107}{20} \Rightarrow b_{XY} = \frac{9}{20} \qquad \therefore \quad r^2 = b_{YX} \cdot b_{XY} = \frac{4}{5} \times \frac{9}{20} = 0.36$$

$$\Rightarrow r = \pm\sqrt{0.36} = \pm 0.6$$

Since both the regression coefficients are positive, r must be positive. Hence, we take $r = r_{XY} = 0.6$

We are given, $\sigma_X = 3$

We have, $b_{YX} = r\frac{\sigma_Y}{\sigma_X} \Rightarrow \sigma_Y = \frac{b_{YX} \cdot \sigma_X}{r} = \frac{4}{5} \times \frac{3}{0.6} = 4$

Correlation Coefficient

The correlation coefficient is the under-root of the two regression coefficients, i.e.,

$r = \sqrt{b_{XY} \times b_{YX}}$

$b_{XY} = 0.95$; $b_{YX} = 0.95$

Hence $r = \sqrt{0.95 \times 0.95} = 0.95$

5.1.9. Standard Error in Regression Analysis

In regression analysis dependent variables are calculated by regression equation while value of independent variable is given. It is almost impossible to calculate 100% accurate estimate. To check the accuracy and reliability one calculates the standard error of estimate. Standard error is calculated to find out that what extent the estimated values are closer to actual values.

There are two standard error of estimate one is for regression of X on Y and second is regression of Y on X:

1) **Standard Error of Estimate of Y on X (S_{yx}):** Following formula is used to calculate the standard error of estimate of Y on X (denoted by S_{yx}).

   i) **First Formula**

   $$S_{yx} = \sqrt{\frac{\Sigma(Y - Y_e)^2}{N}}$$

   Where, Y = actual values; Y_e = estimated values

   ii) **Second Formula**

   $$S_{yx} = \sqrt{\frac{\Sigma Y^2 - a\Sigma Y - b\Sigma XY}{N}}$$

   Where, a and b are to be obtained from normal equations and a = intercept, b = slope of line.

Scanned with CamScanner

iii) **Third Formula**

$$S_{yx} = \sigma_y \sqrt{1-r^2}$$

Where, σ_y = SD of Y
r = coefficient of correlation between X and Y

The above formula is suitable when it is given the values of correlation coefficient (r) and standard deviations (σ_x and σ_y).

2) **Standard Error of Estimate of X on Y (S_{xy}):** It is calculate by following formula and represented by S_{xy}:

i) **First Formula**

$$S_{xy} = \sqrt{\frac{\Sigma(X-X_c)^2}{N}}$$

Where, X = actual values, X_c = estimated values

ii) **Second Formula**

$$S_{xy} = \sqrt{\frac{\Sigma X^2 - a\Sigma X - b\Sigma XY}{N}}$$

Where, a and b are to be obtained from normal equations and a = intercept, b = slope of line.

iii) **Third Formula**

$$S_{xy} = \sigma_x \sqrt{1-r^2}$$

Where, σ_x = SD of Y
r = coefficient of correlation between X and Y.

The above formula is suitable when it is given the values of correlation coefficient (r) and standard deviations (σ_x and σ_y).

Example 5: Find the 'standard error of the estimates':
$\sigma_x = 4.4$, $\sigma_y = 2.2$, r = 0.8

Solution: Given, r =0.8, $\sigma_x = 4.4$, $\sigma_y = 2.2$

We can find out the 'standard error of estimates' as follows:

$$S_{yx} = \sigma_y \sqrt{1-r^2} \quad \text{..... (1)}$$

$$S_{xy} = \sigma_x \sqrt{1-r^2} \quad \text{..... (2)}$$

Putting the given values in equation (1) and (2), we get,

$$S_{yx} = 2.2 \times \sqrt{1-(0.8)^2} = 2.2 \times \sqrt{1-0.64} = 2.2 \times \sqrt{0.36} = 2.2 \times 0.6 = 1.32$$

$$S_{xy} = 4.4 \times \sqrt{1-(0.8)^2} = 4.4 \times \sqrt{1-0.64}$$
$$= 4.4 \times \sqrt{0.36} = 4.4 \times 0.6 = 2.64$$

Example 6: Given, $\sigma_X = 3$, $\sigma_Y = 4$, r = 0.80. Determine the standard error of estimate S_{xy} and S_{yx}.

Solution: Since,

$$S_{xy} = \sigma_X \sqrt{1-r^2} = 3 \times \sqrt{[1-(0.80)^2]} = 3 \times \sqrt{(0.36)} = 1.8$$

$$S_{yx} = \sigma_Y \sqrt{1-r^2} = 4 \times \sqrt{[1-(0.80)^2]} = 4 \times \sqrt{(0.36)} = 2.4$$

Example 7: The expert of hospital administration told that number of FTFs (full-time employees) of hospital can be determined by calculating the total number of beds in the hospital. The business researcher of healthcare decided to make a regression model to calculate the number of FTEs of a hospital by the total number of beds. She took the data of 12 hospitals as shown in **table below.** Data are obtained in sequence, according to total number of beds.

No. of Beds	23	29	29	35	42	46	50	54	64	66	76	78
FTEs	69	95	102	118	126	125	138	178	156	184	176	225

Calculate the sum of squares of error and the standard error of the estimate.

Solution: Let us first determine the value of Σx, Σy, Σx2 and Σxy using the following table:

Hospital	Number of Beds x	FTEs y	x^2	xy
1	23	69	529	1587
2	29	95	841	2755
3	29	102	841	2958
4	35	118	1225	4130
5	42	126	1764	5292
6	46	125	2116	5750
7	50	138	2500	6900
8	54	178	2916	9612
9	64	156	4096	9984
10	66	184	4356	12144
11	76	176	5776	13376
12	78	225	6084	17550
	Σx = 592	Σy = 1692	Σx^2 = 33044	Σxy = 92038

By above values, the healthcare researcher solved for the sample slope(b) and the sample y intercept(a):

$$SS_{xy} = \Sigma xy - \frac{(\Sigma x)(\Sigma y)}{n} = 92038 - \frac{(592)(1692)}{12} = 8566$$

$$SS_{xx} = \Sigma x^2 - \frac{(\Sigma x)^2}{n} = 33044 - \frac{(592)^2}{12} = 3838.667$$

$$b = \frac{SS_{xy}}{SS_{xx}} = \frac{8566}{3838.667} = 2.232$$

$$a = \frac{\Sigma y}{n} - b\frac{\Sigma x}{n} = \frac{1692}{12} - (2.232)\frac{592}{12} = 30.888$$

A Least square equation of the regression line is as follows:

$\hat{y} = 30.888 + 2.232x$.

Slope is b=2.232, this shows that for all unit increase of x(every bed), y(number of FTEs) is projected to increase by value 2.232.

Scanned with CamScanner

Hospital	Number of Beds x	FTES y	Residuals y − ŷ	(y − ŷ)²
1	23	69	−13.22	174.77
2	29	95	−0.62	−0.38
3	29	102	6.38	40.70
4	35	118	8.99	80.82
5	42	126	1.37	1.88
6	46	125	−8.56	73.27
7	50	138	−4.49	20.16
8	54	178	26.58	706.50
9	64	156	−17.74	314.71
10	66	184	5.80	33.64
11	76	176	−24.52	601.23
12	78	225	20.2	400.80
	Σx = 592	Σy = 1692	Σ(y − ŷ) = −.01	Σ(y − ŷ)² = 2448.86

Error Sum of Square (SSE) = 2448.86

Standard Error of Estimate, $s_e = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{2448.86}{10}} = 15.65$

The 15.65 FTEs is the estimated standard error. The analysis of the residuals for this problem tells that eight of 12(67%) are within $\pm 1s_e$ and 100% are within $\pm 2s_e$.

Example 8: Consider the following data and find out the standard error of the estimate of X on Y and Y on X:

X:	1	2	3	4	5
Y:	6	8	7	6	8

Solution:

X	$(X - \bar{X})$ x	x^2	Y	$(Y - \bar{Y})$ y	y^2	xy
1	−2	4	6	−1	1	2
2	−1	1	8	+1	1	−1
3	0	0	7	0	0	0
4	1	1	6	−1	1	−1
5	2	4	8	+1	1	2
ΣX = 15		$\Sigma x^2 = 10$	ΣY = 35		$\Sigma y^2 = 4$	Σxy = 2

$$\bar{X} = \frac{15}{5} = 3,\ \bar{Y} = \frac{35}{5} = 7$$

$$\sigma_x = \sqrt{\frac{\Sigma x^2}{N}} = \sqrt{\frac{10}{5}} = 1.414 \text{ and } \sigma_y = \sqrt{\frac{\Sigma y^2}{N}} = \sqrt{\frac{4}{5}} = 0.894$$

$$r = \frac{\Sigma xy}{N \cdot \sigma_x \cdot \sigma_y}$$

$$r = \frac{2}{5 \times 1.414 \times 0.894} = \frac{2}{6.321} = 0.316$$

$$S_{xy} = \sigma_x \sqrt{1 - r^2} = 1.414\sqrt{1 - (0.316)^2} = 1.414 \times 0.949 = 1.342$$

$$S_{yx} = \sigma_y \sqrt{1 - r^2} = 0.894\sqrt{1 - (0.316)^2} = 0.894 \times 0.949 = 0.848$$

5.2. MULTIPLE REGRESSION

5.2.1. Introduction

Three or more variables, such as X_1, X_2, and X_3, are analysed in multiple regression analysis. Here, X_1 is taken as a dependent variable and its relative movement is found out, according to the movement in both independent variables X_2 and X_3. Thus, the effect of two or more independent variables on one dependent variable is studied in the 'multiple regression analysis'. The procedure for studying multiple regression is similar to simple regression, but there is a single difference that the independent variable is more than one. If there are three variables, X_1, X_2 and X_3, the multiple regression will take the following form:

$$X_1 = a_{1.23} + b_{12.3} X_2 + b_{13.2} X_3 \quad \ldots\ldots (1)$$

In the above equation there are three constraints. The subscript after the dot indicates the variables, which are held constant.

If X_2 and X_3 are 0 then $X_1 = a_{1.23}$

For multiple regression models with two independent variables the result is determined with the use of three simultaneous equations, with three unknown (b_0, b_1 and b_2) values. Statistical software, such as SPSS, MINITAB, MS-EXCEL etc., are used in multiple regression models because this analysis involves lengthy calculations.

5.2.2. Assumptions of Multiple Regression Analysis

Following are the assumptions of multiple regression analysis:

1) The dependent variable should be a random variable whereas the independent variable need not be random variable,
2) The relationship between various independent variables and dependent variable is linear, and
3) The variances of the conditional distributions of the dependent variable and various combinations of values of the independent variables are all equal. The conditional distributions for the dependent variable follows the normal probability distribution.

5.2.3. Steps in Multiple Regression

Step 1) State the research hypothesis.
Step 2) State the null hypothesis.
Step 3) Collect the data.
Step 4) Calculate each variable separately (calculate measures of central tendency and dispersion, frequency distributions, graphs). Check, whether the variables are normally distributed or not.
Step 5) Calculate the relationship between the dependent variable and the independent variables (find the correlation coefficient, obtain a scatter plot). Check, whether the variables are linearly correlated or not.
Step 6) Calculate the relationship between all the independent variables with each other (calculate the correlation coefficient matrix for all the independent variables).
Step 7) Calculate the regression equation from the data.
Step 8) Calculate and examine appropriate measures of association and test of statistical significance for each coefficient as well as for the equation as a whole.
Step 9) Accept or reject the null hypothesis.
Step 10) Reject or accept the research hypothesis.
Step 11) Explain the practical implications of the findings.

Scanned with CamScanner

5.2.4. Difference between Simple and Multiple Regressions

The difference between simple and multiple regressions are shown in below **table 5.2**:

Table 5.2: Comparison between Simple and Multiple Regressions

Simple Regressions	Multiple Regressions
An dependent variable Y predicted from an independent variable, X	An dependent variable, Y predicted from the independent variables set $(X_1, X_2, \ldots, X_k)$
Simple Regression has only one regression coefficient	For each independent variable, there is a regression coefficient.
r^2 is the proportion of variation in dependent variable Y predicted from X.	R^2 is the proportion of variation in dependent variable Y predicted by the independent variables set of X's.

5.2.5. Methods of Calculation

1) **Multiple Regression Equations:** In case of simple regression, there is a regression line approximating a set of N points (X and Y) in a two-dimensional scatter diagram. Similarly, in case of multiple regression there exists a least square regression plane fitting a set of N points (X_1, X_2 and X_3) in a three dimensional scatter diagram:

i) The regression plane of X_1 on X_2 and X_3 is represented by the equation (1),

$$X_1 = a_{1.23} + b_{12.3} X_2 + b_{13.2} X_3 \quad \ldots\ldots (1)$$

The values of $b_{12.3}$ and $b_{13.2}$ are determined by solving simultaneously the following three normal equations:

$$\Sigma X_1 = Na_{1.23} + b_{13.2} \Sigma X_2 + b_{13.2} \Sigma X_3$$
$$\Sigma X_1 X_2 = a_{1.23} \Sigma X_2 + b_{12.3} \Sigma X^2_2 + b_{13.2} \Sigma X_2 X_3$$
$$\Sigma X_1 X_3 = a_{1.23} \Sigma X_3 + b_{12.3} \Sigma X_2 X_3 + b_{13.2} \Sigma X^2_3$$

On multiplying X_2 and X_3 on both sides of equation (1), the last two normal equations are obtained.

ii) The regression equation of X_2 on X_1 and X_3 would be as follows:

$$X_2 = a_{2.13} + b_{21.3} X_1 + b_{23.1} X_3 \quad \ldots\ldots (2)$$

The normal equations for fitting the above equation are:

$$\Sigma X_2 = Na_{2.13} + b_{21.3} \Sigma X_1 + b_{23.1} \Sigma X_3$$
$$\Sigma X_1 X_2 = a_{2.13} \Sigma X_1 + b_{21.3} \Sigma X_1^2 + b_{23.1} \Sigma X_1 X_3$$
$$\Sigma X_2 X_3 = a_{2.13} \Sigma X_3 + b_{21.3} \Sigma X_1 X_3 + b_{23.1} \Sigma X^2_3$$

iii) The regression equation of X_3 on X_1 and X_2 would be as follows:

$$X_3 = a_{3.12} + b_{31.2} X_1 + b_{32.1} X_2 \quad \ldots\ldots (3)$$

The normal equations for fitting the above equation are:

$$\Sigma X_3 = Na_{3.12} + b_{31.2} \Sigma X_1 + b_{32.1} \Sigma X_2$$
$$\Sigma X_1 X_3 = a_{3.12} \Sigma X_1 + b_{31.2} \Sigma X_1^2 + b_{32.1} \Sigma X_1 X_2$$
$$\Sigma X_2 X_3 = a_{3.12} \Sigma X_2 + b_{31.2} \Sigma X_1 X_2 + b_{32.1} \Sigma X^2_2$$

Generalisation for more than three variables is as follows:

$$X_1 = a_{1.234} + b_{12.34} X_2 + b_{13.24} X_3 + b_{14.23} X_4$$

The four normal equations are determined by multiplying the above equation on both sides by X_1, X_2, X_3 and X_4 to find the values of $a_{1.234}$, $b_{13.24}$, $b_{12.34}$, and $b_{14.23}$.

2) **Deviations from the Mean:** If the deviations from the respective means of the various variables are taken into account, then the calculations for solving these equations can be reduced. Thus, if $X_1 = (X_1 - \overline{X}_1)$,

$X_2 = (X_2 - \overline{X}_2)$ and $X_3 = (X_3 - \overline{X}_3)$, the multiple regression equation can be expressed as follows:

$$X_1 = b_{12.3} X_2 + b_{13.2} X_3$$

Following two normal equations are required to be solved, for determining the values of $b_{12.3}$ and $b_{13.2}$:

$$\Sigma X_1 X_2 = b_{12.3} \sum X_2^2 + b_{13.2} \Sigma X_2 \Sigma X_3$$
$$\Sigma X_1 X_3 = b_{12.3} \Sigma X_2 X_3 + b_{13.2} \sum X_2^2$$

The values of $b_{12.3}$ and $b_{13.2}$ can also be determined as follows:

$$b_{12.3} = r_{12.3} \times \frac{\sigma_{1.23}}{\sigma_{2.13}}; \qquad b_{13.2} = r_{13.2} \times \frac{\sigma_{13.2}}{\sigma_{3.12}}$$

Similar calculations are done for the regression equation of X_2 on X_1 and X_3 and for regression equation of X_3 on X_1 and X_2.

3) **Multiple Regression Equation of X_1 on X_2 and X_3:** It can be expressed as follows:

$$X - \overline{X}_1 = \frac{\sigma_1}{\sigma_2}\left[\frac{r_{12} - r_{13}\, r_{23}}{1 - r_{23}^2}\right](X_2 - \overline{X}_2) + \frac{\sigma_1}{\sigma_3}\left[\frac{r_{13} - r_{12}\, r_{23}}{1 - r_{23}^2}\right](X_3 - \overline{X}_3)$$

Similarly, the multiple regression equation of X_2 on X_3 and X_1 can be expressed as follows:

$$X_2 - \overline{X}_2 = \frac{\sigma_2}{\sigma_3}\left[\frac{r_{23} - r_{12}\, r_{13}}{1 - r_{13}^2}\right](X_3 - \overline{X}_3) + \frac{\sigma_2}{\sigma_1}\left[\frac{r_{12} - r_{23}\, r_{31}}{1 - r_{31}^2}\right](X_1 - \overline{X}_1)$$

Lastly, the multiple regression equation of X_3 on X_1 and X_2 would be expressed as follows:

$$X_3 - \overline{X}_3 = \frac{\sigma_3}{\sigma_2}\left[\frac{r_{23} - r_{12}\, r_{13}}{1 - r_{12}^2}\right](X_2 - \overline{X}_2) + \frac{\sigma_3}{\sigma_1}\left[\frac{r_{13} - r_{23}\, r_{12}}{1 - r_{12}^2}\right](X_1 - \overline{X}_1)$$

In this method, determination of regression equations is simple in comparison when the normal equations are required to be solved for determining the regression equations. Here we only need the following values.

$X_1, X_2, X_3, \sigma_1, \sigma_2, \sigma_3$ and r_{12}, r_{13}, r_{23}.

Example 9: Find the multiple regression equation of X_1 on X_2 and X_3 from the data relating to three variables given below:

X_1	4	6	7	9	13	15
X_2	15	12	8	6	4	3
X_3	30	24	20	14	10	4

Solution: The regression equation of X_1 on X_2 and X_3 is

$$X_1 = a_{1.23} + b_{12.3} X_2 + b_{13.2} X_3$$

The three normal equations to get the values of $a_{1.23}$, $b_{12.3}$ and $b_{13.2}$ are

$$\Sigma X_1 = Na_{1.23} + b_{12.3} \Sigma X_2 + b_{13.2} \Sigma X_3$$
$$\Sigma X_1 X_2 = a_{1.23} \Sigma X_2 + b_{12.3} \Sigma X_2^2 + b_{13.2} \Sigma X_2 X_3$$
$$\Sigma X_1 X_3 = a_{1.13} \Sigma X_3 + b_{12.3} \Sigma X_2 X_3 + b_{13.2} \Sigma X^2_3$$

Scanned with CamScanner

X_1	X_2	X_3	X_1X_2	X_1X_3	X_2X_3	X_2^2	X_3^2
4	15	30	60	120	450	225	900
6	12	24	72	144	288	144	576
7	8	20	56	140	160	64	400
9	6	14	54	126	84	36	196
13	4	10	52	130	40	16	100
15	3	4	45	60	12	9	16
$\Sigma X_1 = 54$	$\Sigma X_2 = 48$	$\Sigma X_3 = 102$	$\Sigma X_1X_2 = 339$	$\Sigma X_1X_3 = 720$	$\Sigma X_2X_3 = 1,034$	$\Sigma X_2^2 = 494$	$\Sigma X_3^2 = 2,188$

Substituting the values in the normal equations:

$6a_{1.23} + 48b_{12.3} + 102b_{13.2} = 54$ ..... (1)

$48a_{1.23} + 494b_{12.3} + 1034b_{13.2} = 339$ ..... (2)

$102a_{1.23} + 1034b_{12.3} + 2188b_{13.2} = 720$

Multiplying (1) by 8, we get

$48a_{1.23} + 384b_{12.3} + 816b_{13.2} = 432$ ..... (4)

Subtracting (2) from (4), we get

$110\ b_{12.3} + 218b_{13.2} = -93$ ..... (5)

Multiple (1) by 17, we get

$102\ a_{1.23} + 816b_{12.3} + 1734b_{13.2} = 918$ ..... (6)

Subtracting (3) from (6), we get

$218\ b_{12.3} + 454b_{13.2} = -198$ ..... (7)

Multiplying (5) by 109, we obtain

$11990\ b_{12.3} + 23762\ b_{13.2} = -10137$ ..... (8)

Multiplying (7) by 55, we get

$11990\ b_{12.3} + 24970b_{13.2} = -10890$ ..... (9)

Subtracting (8) from (9), we get

$$1208b_{13.2} = -753 \text{ or } b_{13.2} = \frac{753}{1208} = -0.623$$

Substituting this value of $b_{13.2}$ in equation (5), we get

$110\ b_{12.3} + 218\ (-0.623) = -93; \quad 110\ b_{12.3} = 135.814 - 93$

$$b_{12.3} = \frac{42.814}{110} = +0.389$$

Substituting the values of $b_{12.3}$ and $b_{13.2}$ in equation (1), we get

$6a_{1.23} + 48(0.389) + 102\ (-0.623) = 54$

$6a_{1.23} = 54 + 63.546 - 18.672 = 98.874$

$a_{1.23} = 16.479$

Thus the required regression equation is: $X_1 = 16.479 + 0.389\ X_2 - 0.623\ X_3$.

5.3. PHARMACEUTICAL EXAMPLES

Example 10: Find equation of regression line Y on X for the following data:

X	25	23	28	26	29	21	30	24	33	36
Y	28	26	28	24	29	26	31	32	36	27

Also estimate Y, if X = 55.

Solution:

Calculation of Regression Equations

X	Y	XY	X^2
25	28	700	625
23	26	598	529
28	28	784	784
26	24	624	676
29	29	841	841
21	26	546	441
30	31	930	900
24	32	768	576
33	36	1188	1089
36	27	972	1296
$\Sigma X = 275$	$\Sigma Y = 287$	$\Sigma XY = 7951$	$\Sigma X^2 = 7757$

Regression equation of Y on X is: $Y = a + bX$

To determine the value of a and b, the following two normal equations are to be solved:

$\Sigma Y = Na + b\Sigma X$ .....(1)

$\Sigma XY = a\Sigma X + b\Sigma X^2$ .....(2)

Substituting the values,

$287 = 10a + 275b$ .....(3)

$7951 = 275a + 7757b$ .....(4)

Multiply 275 in equation (3) and 10 in equation (4),

$78925 = 2750a + 75625b$ .....(5)

$79510 = 2750a + 77570b$ .....(6)

Subtract equation (5) from equation (6),

$585 = 1945b$ Or, $b = 0.3$

Putting the value of b in equation (3),

$287 = 10a + 275 \times 0.3 \Rightarrow 287 = 10a + 82.5 \Rightarrow 10a = 204.5 \Rightarrow a = 20.45$

So, regression equation of Y on X is:

$Y = 20.45 + 0.3X$

When, $X = 55$ then, $Y = 20.45 + 0.3 \times 55 = 36.95$

Example 11: Obtain the lines of regression and using regression coefficients find correlation coefficient from the following data:

Year of Service (in years)	5	9	15	19	26	31	35
Monthly Income (₹ in thousand)	22	26	35	39	48	59	65

Solution:

Calculation of Line of Regression

Year of Service (X)	Monthly Income (Y)	$x = (X - \bar{X})$	$y = (Y - \bar{Y})$	x^2	y^2	xy
5	22	−15	−20	225	400	300
9	26	−11	−16	121	256	176
15	35	−5	−7	25	49	35
19	39	−1	−3	1	9	3
26	48	6	6	36	36	36
31	59	11	17	121	289	187
35	65	15	23	225	529	345
$\Sigma X = 140$	$\Sigma Y = 294$	$\Sigma x = 0$	$\Sigma y = 0$	$\Sigma x^2 = 754$	$\Sigma y^2 = 1568$	$\Sigma xy = 1082$

Scanned with CamScanner

$$\overline{X} = \frac{140}{7} = 20 \quad \text{and} \quad \overline{Y} = \frac{294}{7} = 42$$

$$b_{XY} = \frac{\sum xy}{\sum y^2} = \frac{1082}{1568} = 0.69; \quad b_{YX} = \frac{\sum xy}{\sum x^2} = \frac{1082}{754} = 1.43$$

Regression equation of X on Y is given by

$X - \overline{X} = b_{XY}(Y - \overline{Y})$

$\Rightarrow X - 20 = 0.69(Y - 42)$

$\Rightarrow X = 0.69Y - 42 \times 0.69 + 20$

$\Rightarrow X = 0.69Y - 28.98 + 20$

$\Rightarrow X = -8.98 + 0.69Y$

Regression equation of Y on X is given by

$Y - \overline{Y} = b_{YX}(X - \overline{X})$

$\Rightarrow Y - 42 = 1.43(X - 20)$

$\Rightarrow Y = 1.43X - 20 \times 1.43 + 42$

$\Rightarrow Y = 1.43X - 28.6 + 42$

$\Rightarrow Y = 13.4 + 1.43X$

Correlation coefficient $(r) = \sqrt{b_{XY} \times b_{YX}} = \sqrt{0.69 \times 1.43} = \sqrt{0.9867} = 0.993$

Example 12: You are given the following information about advertising expenditures and the sales:

	Advertisement (₹ in Lac)	Sales (₹ in Lac)
Arithmetic Mean	10	90
Standard Deviation	3	12

The coefficient of correlation is 0.8.

i) Obtain two regression equations.

ii) Find the likely sales when the advertisement budget is ₹15 lac.

iii) What should be the advertisement budget if the company wants to attain sales target ₹120 lac?

Solution: Given,

$\overline{X} = 10, \sigma_x = 3, \overline{Y} = 90, \sigma_y = 12$ and $r = 0.8$

$$b_{xy} = r\frac{\sigma_x}{\sigma_y} = 0.8 \times \frac{3}{12} = 0.2; \qquad b_{yx} = r\frac{\sigma_y}{\sigma_x} = 0.8 \times \frac{12}{3} = 3.2$$

i) Regression equation of X on Y is given by

$X - \overline{X} = b_{xy}(Y - \overline{Y})$ Or $X - 10 = 0.2(Y - 90)$

$X - 10 = 0.2Y - 18$ Or $X = 0.2Y - 18 + 10$

$X = 0.2Y - 8$

Regression equation of Y on X is given by

$Y - \overline{Y} = b_{yx}(X - \overline{X})$

$Y - 90 = 3.2(X - 10)$ Or $Y = 3.2X - 32 + 90$ Or $Y = 3.2X + 58$

ii) Likely sales when the advertisement budget is ₹15 lac i.e., $X = 15$

$\Rightarrow \quad Y = 3.2X + 58 \Rightarrow Y = 3.2 \times 15 + 58 \Rightarrow \quad Y = 3.2 \times 15 + 58$

$\Rightarrow \quad Y = 48 + 58 \Rightarrow \quad Y = 106$ lac

iii) Value of advertisement budget when company wants to attain sales target ₹120 lac i.e., $Y = 120$

$\Rightarrow \quad X = 0.2Y - 8 \Rightarrow \quad X = 0.2 \times 120 - 8$

$\Rightarrow \quad X = 24 - 8 \Rightarrow \quad X = 16$ lac

Example 13: Find the mean of X and Y variables and the coefficient of correlation between them form the following two regression equations:

$2Y - X = 50, \; 3Y - 2X = 10$

Solution:

i) The given equations are:

$2Y - X = 50$ ..... (1)

$3Y - 2X = 10$ ..... (2)

Since the mean of X and Y lie on the regression line. So,

$2\overline{Y} - \overline{X} = 50$ ..... (3)

$3\overline{Y} - 2\overline{X} = 10$ ..... (4)

Multiply 2 in equation (1),

$4\overline{Y} - 2\overline{X} = 100$ ..... (5)

Subtract equation (4) from (5),

$\overline{Y} = 90$

Putting the value of $\overline{Y}$ in equation (3),

$2 \times 90 - \overline{X} = 50$ Or $180 - \overline{X} = 50$ Or $\overline{X} = 130$

Hence means are,

$\overline{X} = 130 \quad \overline{Y} = 90.$

ii) Let us assume that the first equation be regression equation of Y on X.

$$2Y - X = 50 \text{ Or } Y = \frac{X}{2} - 25 \quad \therefore \quad b_{yx} = \frac{1}{2}$$

Second equation will be regression equation of X on Y.

$3Y - 2X = 10; \quad 2X = 3Y - 10$

$$X = \frac{3}{2}Y - 5 \quad \therefore \quad b_{xy} = \frac{3}{2}$$

Now, $$r = \sqrt{b_{xy} \cdot b_{yx}} = \sqrt{\frac{3}{2} \times \frac{1}{2}} = \sqrt{\frac{3}{4}} = \sqrt{0.75} = 0.866$$

Example 14: Write regression equations of X on Y and of Y on X for the following data:

X	45	48	50	55	65	70	75	72	80	85
Y	25	30	35	30	40	50	45	55	60	65

Solution: We prepare the table for working out the values for the regression lines.

X	Y	$U = X - 65$	$V = Y - 45$	U^2	V^2	UV
45	25	-20	-20	400	400	400
48	30	-17	-15	289	225	255
50	35	-15	-10	255	100	150
55	30	-10	-15	100	225	150
65	40	0	-5	0	25	0
70	50	5	5	25	25	25
75	45	10	0	100	0	0
72	55	7	5	49	25	35
80	60	15	15	225	225	225
85	65	20	20	400	400	400
$\Sigma X = 645$	$\Sigma Y = 435$	$\Sigma U = -5$	$\Sigma V = -15$	$\Sigma U^2 = 1813$	$\Sigma V^2 = 1725$	$\Sigma UV = 1675$

Scanned with CamScanner

We have,

$$\bar{X} = \frac{\sum X}{N} = \frac{645}{10} = 64.5 \text{ and } \bar{Y} = \frac{\sum Y}{N} = \frac{435}{10} = 43.5$$

$$b_{yx} = \frac{N\sum UV - (\sum U \sum V)}{N\sum U^2 - (\sum U)^2} = \frac{(10)x1415 - (-5)x(-15)}{(10)x1813 - (-5)^2} = \frac{14150 - 75}{18130 - 25} = \frac{14075}{18105} = 0.777$$

Regression equation of Y on X is:

$Y - \bar{Y} = b_{yx}(X - \bar{X})$ $Y - 43.5 = 0.777\ (X - 64.5)$ Or $Y = 0.787X + 7.26$

Similarly b_{xy} can be calculated as:

$$b_{xy} = \frac{N\sum UV - (\sum U \sum V)}{N\sum V^2 - (\sum V)^2} = \frac{(10)x1415 - (-5)x(-15)}{(10)x1675 - (-15)^2} = \frac{14150 - 75}{16750 - 225} = \frac{14075}{16525} = 0.85$$

Regression equation of X on Y will be:

$X - \bar{X} = b_{xy}(Y - \bar{Y})$ Or $X - 64.5 = 0.85\ (Y - 43.5)$ Or $X = 0.85\ Y + 26.65$

Example 15: The data of ages and blood pressure of 10 women are shown in table below:

Age (X)	56	42	36	47	49	42	60	72	63	55
Blood Pressure (Y)	147	125	118	128	145	140	155	160	149	150

1) Find the least square regression equation of Y on X.
2) Calculate the blood pressure of 45 years old woman.

Solution:

1) Let $y = a + bx$ be straight line equation.
Normal equations are as follows:

$$\sum y = an + b\sum x; \qquad \sum xy = a\sum x + b\sum x^2$$

The following table shows the calculation of values:

x	y	x^2	x·y
56	147	3136	8232
42	125	1764	5250
36	118	1296	4248
47	128	2209	6016
49	145	2401	7105
42	140	1764	5880
60	155	3600	9300
72	160	5184	11520
63	149	3969	9387
55	150	3025	8250
–	–	–	–
$\sum x = 522$	$\sum y = 1417$	$\sum x^2 = 28348$	$\sum x \cdot y = 75188$

Putting all these values in normal equations:
$10a + 522b = 1417$
$522a + 28348b = 75188$

By using elimination method, we can solve both equations,
$10a + 522b = 1417$, and
$522a + 28348b = 75188$
Or, $2(261a + 14174b) = 2 \times 37594$ Or, $261a + 14174b = 37594$

Thus, we have two equations:
$10a + 522b = 1417$ .....(1)
$261a + 14174b = 37594$ .....(2)
Equation (1) $\times 261 \Rightarrow 2610a + 136242b = 369837$
Equation (2) $\times 10 \Rightarrow 2610a + 141740b = 375940$

Subtracting above equations, we get:

$$-5498b = -6103 \Rightarrow 5498b = 6103 \Rightarrow b = \frac{6103}{5498} \Rightarrow b = 1.11$$

Substituting b= 1.11 in equation (1), we get:
$10a + 522\ (1.11) = 1417$
$\Rightarrow 10a = 1417 - 579.4409$
$\Rightarrow 10a = 837.5591 \Rightarrow a = \frac{837.5591}{10} \Rightarrow a = 83.7559 \therefore a = 83.7559$ and $b = 1.11$

So putting the values of a and b in equation $y = a + bx$, we have:
$y = 83.7559 + 1.11x$

2) For a women of 45 years old, the calculated average blood pressure will be as follows:
$y = 83758 + 111x$ Or, $y = 83.758 + 1.11\ (45)$ Or, $y = 83.758 + 49.95$
Or, $y = 133.708$

So the blood pressure of 45 years old woman is 134.

Example 16: The Beer's law plot is built by plotting ultraviolet absorbance *vs.* concentration. The result is shown in table below:

Concentration (x)	1	2	3	5	10
Absorbance (y)	0.10	0.36	0.57	1.09	2.05
xy	0.10	0.72	1.71	5.14	20.50

1) Determine the slope and intercept
2) 1.65 is the absorbance of an unknown. Find the concentration?

Solution: Here N =5

x	y	xy	x^2	$(x - \bar{x})$	$(y - \bar{y})$	y^2
1	0.10	0.10	1	–3.2	–0.734	0.01
2	0.36	0.72	4	–2.2	–0.474	0.1296
3	0.57	1.71	9	–1.2	–0.264	0.3249
5	1.09	5.14	25	0.8	0.256	1.1881
10	2.05	20.50	100	5.8	1.216	4.2025
$\sum x = 21$	$\sum y = 4.17$	$\sum xy = 27.17$	$\sum x^2 = 139$			$\sum y^2 = 5.8551$

1) Let the Regression Line of x on y is as follows:
$x = a + by$ .....(1)

Then the two normal equations are as follows:
$\sum x = N\,a + b\sum y$ ..... (2)

Where N= Total number of observations.
Similarly, $\sum xy = a\sum y + b\sum y^2$ ..... (3)

Substituting these values in equation (2), we have:
$21 = 5a + 4.17b$ ..... (4)
Where N is 5;

Also, substituting values from the table into equation (3), we have:
$27.17 = 4.17a + 5.8551b$ ..... (5)

Scanned with CamScanner

Now multiplying equation (4) by 4.17 and equation (5) by 5, we get:

$+87.57 = +20.85a + 17.38b$

$+135.85 = +20.85a + 29.27b$ ..... (6)

$-48.28 = -11.89b$

$\therefore 11.89b = 48.28$ Or $b = \frac{48.28}{11.89}$ Or $b = 4.06$

Substituting the value of b in in equation (4), we get:

$21 = 5a + 4.17 \times 4.06$

$21 = 5a + 16.93$ Or $5a = 21 - 16.93$ Or $a = \frac{4.07}{5}$ Or $a = 0.814$

So, substituting the values of a and b in equation (1), we get the following equation:

$x = a + by$

$x = 0.814 + 4.06y$

$\therefore 4.06y = x - 0.814$ ..... (7)

2) It is given that $y = 1.65$

Substituting the value of y in equation (7), we get:

$(4.06)(1.65) = x - 0.814$ Or $x = 6.699 + 0.814$ Or $x = 7.513$

Example 17: To study the relationship between the shell height x and length y (in mm) in palelloidapygmaea, an experiment is conducted. In the indo-pacific area, a limpet found attached to rocks and shells along sheltered. Obtained information is given below:

$\Sigma x = 56.6$, $\Sigma y = 151.1$, $\Sigma xy = 311.96$, $\Sigma x^2 = 117.68$, $\Sigma y^2 = 832.85$, $n = 28$

Find the line of regression of y on x.

Solution: By using direct method, this question can be solved. Here we ensure the following two normal equations of regression of y on x i.e., $y = a + bx$.

$\Sigma y = na + b\Sigma x$ .....(1)

$\Sigma xy = a\Sigma x + b\Sigma x^2$ .....(2)

In above normal equations, substituting the given values, we get:

$151.1 = 28a + 56.6b$ .....(3)

$311.96 = 56.6a + 117.58b$ .....(4)

Multiplying equation (3) by 56.6 and equation (4) by 28 and then subtracting equation (4) from equation (3) we have the following:

$8552.26 = 1584.8a + 3203.56b$

$8734.88 = 1584.8a + 3295.04b$

$-182.62 = -91.48b$

$\therefore \quad b = \frac{182.62}{91.48} = 1.99$

In equation (3), substituting the values of 'b', we get:

$151.1 = 28a + 56.6 \times 1.99$ Or, $151.1 = 28a + 112.634$

Or, $28a = 151.1 - 112.634$ Or, $28a = 38.466$

$\therefore \quad a = \frac{38.466}{28} = 1.37$

5.4. EXERCISE

5.4.1. Very Short Answer Type Questions

1) What is regression?
2) What are regression coefficients?
3) Discuss the multiple regressions.
4) What do you mean by simple regression?

5.4.2. Short Answer Type Questions

1) And write the properties of regression coefficients.
2) What do you mean by regression lines and equations?
3) Write the difference between the correlation and regression analysis.
4) Write the application of regression analysis.

5.4.3. Long Answer Type Questions

1) What do you understand by the term 'regression'? With the help of few examples illustrate how regression analysis helps in business decision making.
2) Two regression lines are $3X + 2Y - 26 = 0$ and $6X + Y - 31 = 0$. Find the mean value and the correlation coefficient. If the variance of X is 25 then find S.D. of Y.

   [Ans: Mean of X = 4, mean of Y = 7, r = 0.5, S.D. of Y = 15]
3) From the following data calculate: i) Correlation, ii) S.D. of Y

   $X = 0.854Y$ $Y = 0.89X$ and S.D. $X = 3$

   [Ans: r = 0.87, S.D. of Y = 3.06]
4) From the following data obtain the two regression lines:

X	6	2	10	4	8
Y	9	11	5	8	7

   [Ans: $x + 1.3y = 16.4$, $y + 0.65x = 11.9$]
5) From the following data obtain the two regression lines:

X	1	5	3	2	1	1	7	3
Y	6	1	0	0	1	2	1	5

   i) Predict 'y' when x = 10
   ii) Predict 'x' when y = 2.5

   [Ans: $0.3x + y = 2.09$, $x + 0.28y = 3.56$, (i) 0.10, (ii) 2.86]
6) The following table gives the ages (X) and blood pressure (y) of 10 men:

X	56	42	36	47	49	42	60	72	63	55
Y	147	125	118	128	145	140	155	160	149	150

   i) Obtain regression line Y on X.
   ii) Estimate blood pressure of a man whose age is 45 years.

   [Ans: $Y = 1.11X + 83.75$, 134]
7) Calculate regression coefficients for the following data:

X	2	4	6	8	10	12	14
Y	4	2	5	10	4	11	12

   [Ans: $b_{xy} = 0.85$, $b_{yx} = 0.73$]

Scanned with CamScanner

CHAPTER 6 Probability

6.1. PROBABILITY

6.1.1. Meaning and Definition

Probability is the mode of expression of knowledge or belief whether the event has occurred or will occur. Thus, in the theory of probability this concept has been given an exact mathematical significance for being used widely in such areas of study as mathematics, statistics, finance, gambling, science, and philosophy, so that conclusions can be drawn regarding the probability of potential events and the mechanics of complex systems being hidden.

According to Levin, "We live in a world in which we are unable to forecast the future with complete certainty. Our need to cope with uncertainty leads us to the study and use of probability theory."

According to American Heritage Dictionary, "Probability is the branch of mathematics that studies the likelihood of occurrence of random events in order to predict the behavior of a defined system."

General Rule for Probability
The probability of an event (say A) is defined as follows:

If an event happens m times and it fails n times, then the probability (p) of event A is defined as: $P(A) = \frac{m}{m+n}$

or $P(A) = \frac{\text{Number of cases favorable to the occurrence of the event}}{\text{Total number of mutually exclusive and exhaustive cases}}$

So the total number of trials favorable to the event (A) divided by the total number of ways in which the event (A) can happen.

or $P(A) = \frac{\text{Number of cases favorable to the occurrence of the event}}{\text{Total number of mutually exclusive and exhaustive cases}}$

Probability in favor of the occurrence of the event

$$\Rightarrow \frac{m}{n} = \frac{\text{Number of cases favorable to the occurrence of the event}}{\text{the number of cases against the occurrence of the event}}$$

Probability against the occurrence of the event

$$\Rightarrow \frac{n}{m} = \frac{\text{The number of cases against the occurrence of the event}}{\text{The number of cases favorable to the occurrence of the event}}$$

6.1.2. Applications of Probability

1) It plays a great role in making predictions.
2) It is used to solve scientific investigations as well as day-to-day problems.

3) It is very useful when one wants to predict the uncertainties of betting and chances of success.
4) It is extensively used in business and economics.
5) It is the basis of the law of 'statistical regularity' and 'law of inertia of a large number'.

6.1.3. Characteristics of Probability Function P(A)

1) The limit of probability of an event A is between 0 and 1, i.e. $0 \le P(A) \le 1$.
2) When probability of two events A and B which are mutually exclusive, is combined, the probability is the sum of individual probabilities of A and B, i.e., $P(A \cup B)$ will be $P(A) + P(B)$.
3) Two events which are equivalent, are assigned the same probability.

6.1.4. Basic Concepts of Probability

The different basic concepts of probability are as follows:
1) Experiment 2) Sample space 3) Events

6.1.4.1. Experiment

A statistical experiment is an unsystematic or non-deterministic experiment. Following are the features of an experiment:
1) The experiment can have more than one possible outcome.
2) Each possible outcome can be described in advance.
3) The outcome of the experiment is based on possibility.

Random experiment model (in this model outcome is unpredictable with certainty, before the running of the experiment.) is the basis of the probability theory. It assumes that the experiment is repeated infinite number of times under the same conditions. This assumption is significant because the theory of probability is concerned with the replication of the experiment and also has long term behaviour. In this model the term parameter is a non-random quantity which is chosen once and remains constant. There are one or more parameters in various models of random experiment that are adjusted and fit into the physical experiment which is being modelled.

Examples:
1) In tossing of a fair coin which is a random experiment, there are two possible outcomes, either "head up" or "tail up", but it is not sure which face will turn up.
2) Rolling a fair die, which is a random experiment, having six possible outcomes, i.e., appearance of 1or 2 or 3or 4 or 5 or 6 on the top face of the die.
3) Random experiment of drawing a card from a well-shuffled pack of 52 playing cards. In a specific trial there is a chance of drawing any card amounting to 52 possible outcomes.

6.1.4.2. Sample Space

A sample space is a collection of all possible outcomes of a **random experiment.** Mathematically, the sample space is denoted by the symbol S.

For example, the sample space of a toss of two coins, each of these has heads (H) or tails (T), is the set of all possible outcomes: HH, HT, TH, and TT.

6.1.4.3. Events

Events are the possible outcomes of a trial/experiment which are generally denoted by capital letters A, B, C, etc. **For example,** appearance of the outcomes 1 or 2 or 3 or 4 or 5

Scanned with CamScanner

or 6 are called events on the rolling of a fair die. The sample space S (it always occurs) and null set ϕ (it never occurs) is also an event. If M and N are two events in a trial and M is a subset of N, then the occurrence of M implies the occurrence of N. **For example,** while tossing two coins simultaneously the possible outcomes are {HH, HT, TH, TT}. "Getting two heads" is the one possible outcome of an event from the set of all outcomes. Therefore, HH is a simple event. The event 'getting two of the same side, i.e., HH and TT are two outcomes, is a compound event.

Types of Events

There are various kinds of events in probability. They are defined below:

1) **Simple Events:** A simple event is an event that consists of a single possible outcome of an experiment or one can say that, a simple event is the event of a single outcome. The word simple is used to refer to an event which consists of exactly one outcome. **For example,** in the experiment of flipping two coins at the same time the following events are simple events:

   A: 2 heads appear (HH) B: two tails appears (TT)

2) **Compound Events:** A compound event is an event in which there is more than one possible outcome. **For example,** Tossing three dice is a compound event. There are eight possible outcomes:

   HHH, HHT, HTH, HTT, THH, THT, TTH, TTT

3) **Complementary Events:** Complimentary events are events that cannot occur at the same time. The complement of an event A is the set of all outcomes in the sample space that are not included in the outcomes of event A. The event A and its complement $\overline{A}$ are mutually exclusive and exhaustive. The complementary event $\overline{A}$ to an event A is that 'A does not occur'. It satisfies $A \cup \overline{A} = S$, where S is the sample space, and $A \cap \overline{A} = \phi$, where ϕ is the empty set.

   For example, when a die is thrown:

   i) Getting 'even face' and 'odd face' are complimentary events to each other.
   ii) Getting 'divisible by 3' and getting 'not divisible by 3' are complement events of the latter event.
   iii) Getting 'multiple of 2' and getting 'not multiple of 2' are complement events of the latter event.

4) **Favourable and Unfavourable Events:** In a random experiment the number of outcomes resulting in the occurrence of a particular event is termed to be the case, **favorable** to the event. **For example,** if one throws a six-sided die and defines an event A such that the "outcome of an even number", thus each of the cases 2, 4 and 6 are favourable to the event A. In a random experiment, the number of outcomes resulting in not happening of a particular event is termed as the cases being **unfavourable** of an event. **For example,** if one throws a six-sided die and defines an event A such that the "outcome of an even number", thus each of the cases 1, 3 and 5 are unfavourable to the event A.

   So, unfavourable events can be defined as,

   Unfavourable Events = Total Possible Events – Favourable Events

5) **Mutually Exclusive Events:** Two or more events are said to be mutually exclusive events if it is not possible for them to occur together at the same time. In other words, events that have no outcomes in common are said to be mutually exclusive events or disjoint events. Now, one can say that the two events A and B are mutually exclusive if and only if $A \cap B = \phi$. **For example,** consider the experiment of

throwing a six-sided dice and A be the event of getting an odd number whereas B be the event of getting an even number. Therefore,

$A = \{1, 3, 5\}$ and $B = \{2, 4, 6\}$

The above events A and B are **mutually exclusive**, since the two events cannot occur simultaneously. Mathematically,

$(A \cap B) = \phi$.

6) **Mutually Exhaustive Events:** If at least a particular event occurs necessarily from a set of events then it is called mutually exhaustive events. **For example,** for throwing a six-sided dice the events A: number appears greater than 4 and B: number appears less than or equal to 4 are mutually exhaustive events as one of the events is sure to occur.

7) **Equally likely Events:** Two or more outcomes of an experiment are said to be equally likely if each of them has equal chance of happening. **For example,** consider the experiment of throwing a six sided dice, all possible outcomes 1, 2, 3, 4, 5, 6 are equally likely because none of the outcomes can occur in preference over the other.

8) **Independent Events:** Events are said to be independent, if the occurrence of one of them does not affect the occurrence of any of the other. **For example,** in tossing two coins, the occurrence of head on first coin does not affect the occurrence or occurrence of head on the second coin.

9) **Dependent Events:** Two events are said to be dependent, if the occurrence of one event affects the occurrence of the other event. **For example,** the probability of drawing a queen from a pack of 52 cards is 4/52; the card is not put back; then the probability of drawing a queen again is 3/51. Thus, the outcome of the first event affects the outcome of the second event and they are dependent. But, if the cards are put back, then the probability of drawing a queen is 4/52 and is an independent event.

6.1.5. Theorems of Probability

The following are the two important and basic theorems of probability:

1) Multiplication theorem of probability.
2) Addition theorem of probability, and

6.1.5.1. Multiplication Theorem of Probability

To obtain the probability of combined/simultaneous occurrence of several events, the theorem of multiplication probability is used.

The description of this rule for both independent and dependent events is described as follows:

1) **Multiplication Theorem on Probabilities for Independent Events:** If two events A and B are independent, then the probability of simultaneous occurrence of both the events is equal to the product of their individual properties. Thus, $P(A \cap B) = P(A).P(B)$

2) **Multiplication Theorem on Probabilities for Dependent Events:** If A and B be two dependent events, where the occurrence of one event is affected by the occurrence of the other event, then the probability that both A and B will occur is $P(A \cap B) = P(A)P(B/A)$

Example 1: A coin is tossed three times in succession. What is the probability that all the three are heads?

Solution: Let, A denote the event that first toss is "a head", B denote the event that second toss is "a head", C denote the event that third toss is "a head".

Scanned with CamScanner

Now the result of the first toss does in no way influence the result of the second one and hence A, B, C are all independent events. Moreover,

$$P(A) = P(B) = P(C) = \frac{1}{2}$$

$\therefore$ P (that all the three get heads) = $P(A \cap B \cap C)$

= P(A).P(B).P(C) as all events are independent. $= \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{8}$

Example 2: A person is known to hit a target in 4 out of 5 shots, where as another person is known to hit in 3 out of 4 shots. Find the probability that the target is hit at all when they both try.

Solution: Let A be the event that the first person hits the target,

$$\therefore P(A) = \frac{4}{5} \text{ and } P(A^c) = \frac{1}{5}$$

Let B be the event that another person hits the target,

$$\therefore P(B) = \frac{3}{4} \text{ and } P(B^c) = \frac{1}{4}$$

Let X be the event that the target is hit at all when they both try. This will happen when at least one of them hits the target.

$\therefore$ $X = A \cup B$

Again note that it is easy to calculate the probability of the complement of X and also that A and B are independent events.

$\therefore$ $P(X^c)$ = P (that the target is not hit at all)

$= P(A^c \cap B^c)$

$= P(A^c).\ P(B^c) = \frac{1}{5} \times \frac{1}{4} = \frac{1}{20}$

[Independent events]

$\therefore$ P(X) = P (that the target is hit at all)

$= 1 - P(X^c) = 1 - \frac{1}{20} = \frac{19}{20}$

Example 3: Sexually Transmitted Disease

Assume that there are two doctors, A and B, who tests every patient arriving into a clinic for syphilis. Suppose events A^* = [a positive diagnosis done by doctor A] and B^* = [a positive diagnosis done by doctor B]. Assume doctor A diagnoses 10% of all patients as positive, doctor B diagnoses 17% of all patients as positive, and both doctors diagnose 8% of all patients as positive. Determine that whether the events A^*, B^* are independent or not?

Solution: Given that:

$P(A^*)=0.1$, $P(B^*)=0.17$, $P(A^* \cap B^*) = 0.08$

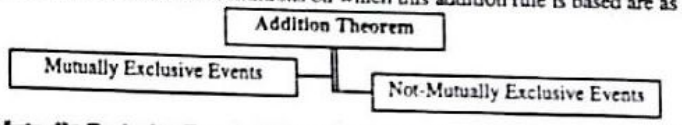
$P(A^*) \times P(B^*) = 0.1(0.17) = 0.017$

Thus, $P(A^* \cap B^*) = 0.08 > 0.017$

This shows that both events are dependent. Because there should be a similarity among how 2 doctors diagnose patients for syphilis, hence this result would be expected.

6.1.5.2. Addition Theorem of Probability

Addition theorem states that one can perform the addition operation between two events which occurs during an experiment and it shows the probability of occurrence of at least one of the several events. The conditions on which this addition rule is based are as follows:



1) **Mutually Exclusive Events:** If there is no sample point common to both the events then they are known as mutually exclusive events. Let us consider that A and B are two mutually exclusive events then $(A \cap B) = \phi$. It means the intersection of two mutually exclusive events is a null set and here $P(A \cap B) = 0$.

$\therefore$ In case of mutually exclusive events,

$$P(A \cup B) = P(A) + P(B)$$

If A, B, C are three mutually exclusive events then,

$$P(A \cup B \cup C) = P(A) + P(B) + P(C)$$

In case of finite number say *n* of mutually exclusive events

$$P(A_1 \cup A_2 \cup A_3 \ldots \cup A_n) = P(A_1) + P(A_2) + \ldots + P(A_n)$$

2) **Not-Mutually Exclusive Events:** If two events A and B are not mutually exclusive events then the probability of occurrence of at least one of them is denoted by $P(A \cup B)$ and is given by,

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Where,

P (A) = Probability of the occurrence of event A.

P (B) = Probability of the occurrence of event B.

$P(A \cap B)$ = Probability of combined occurrence of events A and B.

Let, there are three events A, B and C. The probability of the occurrence of atleast one of them is given by,

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(B \cap C) - P(A \cap C) + P(A \cap B \cap C)$$

The important points to be considered are:

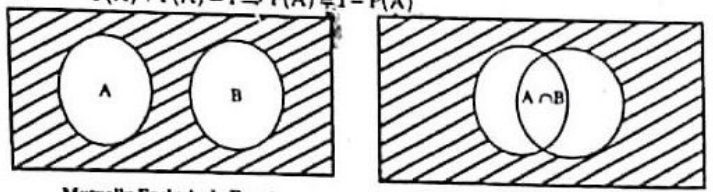
i) If a number of events $A_1, A_2 \ldots A_n$ are mutually exclusive and exhaustive then the sum of the individual probabilities of their happenings is equal to 1, i.e.,

$$P(A_1) + P(A_2) + \ldots + P(A_n) = 1$$

ii) If the events are finite and mutually exclusive then the probability of the occurrence of atleast one of them is equal to the sum of their individual probabilities.

iii) The event A and its compliment $\bar{A}$ are mutually exclusive and exhaustive.

$\therefore$ $P(A) + P(\bar{A}) = 1 \Rightarrow P(\bar{A}) = 1 - P(A)$



Mutually Exclusively Events — Events are Not Mutually Exclusive

Figure 6.1: Two Events Associated with the Addition Theorem

Scanned with CamScanner

Example 4: Find the chance of getting a sum of 8 or more in a single point of throw with two dice.

Solution: The sample space consisting of 36 elements.
Now let A denote the event of throwing a sum 8 in a single throw with two dice. Therefore, A consists of five elements points: (2, 6), (3, 5), (4, 4), (5, 3), (6, 2), and hence $P(A) = \frac{5}{36}$.

Let B denote the event of throwing 9 in a single throw of two dice. Therefore, B consists of 4 sample points:

(3, 6), (4, 5), (5, 4), (6, 3), and hence $P(B) = \frac{4}{36}$.

Let C denote the event of throwing 10 in a single throw of two dice. Therefore, C consists of 3 sample points: (4, 6), (5, 5), (6, 4), and hence $P(C) = \frac{3}{36}$.

Let D denote the event of throwing 11 in single throw of two dice. Therefore, C consists of sample points: (5, 6), (6, 5), and hence $P(D) = \frac{2}{36}$.

Let E denote the event of throwing 12 in a single throw of two dice. Therefore, D consists of only one sample points: (6, 6), and hence $P(E) = \frac{1}{36}$.

Now events A, B, C, D, E are mutually exclusive and hence, we get P(throwing 8 or more in a single throw with two dice),

$= P(A) + P(B) + P(C) + P(D) + P(E)$ [$\because$ Addition rule for disjoint event]

$$= \frac{5}{36} + \frac{4}{36} + \frac{3}{36} + \frac{2}{36} + \frac{1}{36} = \frac{15}{36} = \frac{5}{12}$$

Example 5: A card is drawn from a well-shuffled pack of playing cards. What is the probability that it is either a spade or a king?

Solution: The equi-probable sample space of the experiment of drawing a card from a well shuffled pack of playing cards consists of 52 sample points.
Let A be the event that the card drawn is a spade. Then A consists of 13 sample points then, $P(A) = \frac{13}{52}$.

Let B be the event that the card drawn is a king. Then B consists of 4 kings then $P(B) = \frac{4}{52}$.

Now $(A \cap B)$ consists of only one sample point, the king of spades then $P(A \cap B) = \frac{1}{52}$.

Hence, P (the card drawn is either a spade or a king)

$= P(A \cup B) = P(A) + P(B) - P(A \cap B)$ [By addition rule]

$$= \frac{13}{52} + \frac{4}{52} - \frac{1}{52} = \frac{16}{52} = \frac{4}{13}$$

Example 6: Sexually Transmitted Disease: Let us consider the data given in example 6. Assume that patients are advised for the more lab tests if the doctor A or B makes a positive diagnosis. Find out the probability that patient will be referred for more lab tests?

Solution: The result that either doctor makes a positive diagnosis can be shown by $(A^+ \cup B^+)$. It is given that:
$P(A^+) = 0.1$; $P(B^+) = 0.17$; $P(A^+ \cap B^+) = 0.08$

Hence, by the additional law of probability, we have:
$P(A^+ \cup B^+) = P(A^+) + P(B^+) - P(A^+ \cap B^+) = 0.1 + 0.17 - 0.08 = 0.19$

This shows that 19% of all patients will be referred for more lab tests.

6.1.6. Conditional Probability

Generally, conditional probability analysis involves estimating the probability of occurrence of a particular event (A) as conditional upon the occurrence of a particular event (B). This probability is usually written as P(A/B) and is read "the probability of A given B."

Or

For two events A and B, then the conditional probability of A (given B) is that the probability of occurring A (given that B has occurred) is given by:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad \text{Provided } P(B) > 0.$$

Similarly, the conditional probability of B, given A, denoted by P(B) is defined by:

$$P(B|A) = \frac{P(A \cap B)}{P(A)} \quad \text{Provided } P(A) > 0.$$

For example, consider the experiment of throwing an eight sided dice. There can be eight possible outcomes in this experiment. These eight possible outcomes will form a sample space. So, the sample space will be,

$S = \{1, 2, 3, 4, 5, 6, 7, 8\}$

Let A be the event of getting an odd number, B is the event getting at least 7.

Then, $A = \{1, 3, 5, 7\}$, $B = \{7, 8\}$ and $A \cap B = \{7\}$

$\therefore P(A) = \frac{4}{8} = \frac{1}{2}$, $P(B) = \frac{2}{8} = \frac{1}{4}$ and $P(A \cap B) = \frac{1}{8}$.

$$P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{1/8}{1/2} = \frac{1}{8} \times \frac{2}{1} = \frac{1}{4};$$

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{1/8}{1/4} = \frac{1}{8} \times \frac{4}{1} = \frac{1}{2}$$

Example 7: Cancer
Assume that among 1,00,000 women with negative mammogram there are 20 women will be diagnosed with breast cancer within 2 years, or $P(B|\bar{A}) = 20/10^5 = 0.0002$, while one woman in 10 with positive mammograms will be diagnose with breast cancer within 2 years, or $P(B|A) = 0.1$. The two events A and B would be highly dependent, as RR = $P(B|A)/P(B|\bar{A}) = 0.1/0.0002 = 500$

Scanned with CamScanner

This shows that women with positive mammogram are 500 time more probable to have breast cancer over the upcoming 2 years than women with negative mammograms. This is the basis for using the mammogram as a screening test for breast cancer. If events A and B were independent, then the RR would be 1, in other words, women with positive or negative mammograms would be equally probable to survive with breast cancer. In this case, mammogram would not be useful as a screening test for breast cancer.

Example 8: A pair of dice is rolled. If the sum of 8 has appeared, find the probability that one of the dice shows 3.

Solution: The equi-probable sample space consisting of 36 sample points. The event A = The sum of the scores is 8 has five sample points (2, 6), (3, 5), (4, 4), (5, 3), (6, 2) and it has reduced sample space. $P(A) = \frac{n(A)}{n(S)} = \frac{5}{36}$

Under the assumption that A has happened, the event B = one of the dice shows 3 has only two sample points, that is $(A \cap B) = [(3, 5), (5, 3)]$.

$$\therefore \quad P(A \cap B) = \frac{n(A \cap B)}{n(S)} = \frac{2}{36} = \frac{1}{18}$$

Also using the formula, we get $P(B/A) = \frac{P(A \cap B)}{P(A)} = \frac{1/18}{5/36} = \frac{1}{18} \times \frac{36}{5} = \frac{2}{5}$

Example 9: An box contains 3 white and 5 red balls. Two draws of one ball each are made without replacement. What is the probability that one is red and the other white?

Solution: Probability of drawing a red ball in the first draw, $P(A) = \frac{5}{8}$.

Probability of drawing a white ball in the second draw given that the first draw has given a red ball or $P\left(\frac{B}{A}\right) = \frac{3}{7}$ (since only 7 balls are left in the bag and three white balls are still there).

Probability of the combined event

$$P(AB) = P(A \cap B) = P(A) \times P(B/A) = \frac{5}{8} \times \frac{3}{7} = \frac{15}{56}$$

But it could also happen that in the first draw a white ball was drawn then, Probability of drawing a white ball in the first draw, $P(A) = \frac{3}{8}$, and

Probability of drawing a red ball in the second draw given that the first draw gave a white ball or $P\left(\frac{B}{A}\right) = \frac{5}{7}$.

The combined probability of the two events is,

$$P(AB) \text{ or } P(A \cap B) = P(A) \times P(B|A) = \frac{3}{8} \times \frac{5}{7} = \frac{15}{56}$$

Now any one of the two situations (when we draw a red ball first or we draw a white ball first), would satisfy the conditions of the problem. These two events are mutually exclusive. So the probability that any one of the two happens is the sum of the two probabilities i.e.,

$$\text{Required probability} = \frac{15}{56} + \frac{15}{56} = \frac{15}{28}$$

6.1.7. Baye's Theorem

There are various names assigned to this probability, like, posterior probability, revised probability, or inverse probability. This was introduced by **Thomas Baye** (an English mathematician) in his famous work on Bayesian Decision Theory. This theory states about the finding of the probability of an event by considering the available sample information.

According to **Baye's Theorem** uncertainty can be measured. On the basis of probability theory, this theorem establishes a rule for refining a hypothesis by allotting additional evidence and circumstantial information which results in a number, representing the degree of probability that hypothesis is true. **For example,** for an event A, a chosen sample of 3 defective items out of 100 can be used to estimate the probability for an event B that a machine is not working properly. The Bayesian probability is based on the conditional probability, where A_1 and A_2 are the events that are mutually exclusive and exhaustive. Whereas B is a simple event which intersects with each of the A events as displayed in the Venn diagrams as shown in **figure 6.2**.

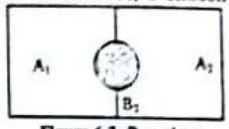


Figure 6.2: Bayesian Probability

This is known as posterior probability as it is calculated after the information is recorded. It is the revised probability as it is derived by the revision of the prior probabilities under the accumulated additional information. Further, it is also known as inverse probability, i.e., determining the probability of a probability. However, the Bayesian, or the Posterior probabilities are always conditional probabilities which are calculated for every events as follows:

1) **Mutually Exclusive Events:** If there exists the occurrence of an event E in combination with one of the mutually exclusive events $E_1, E_2 \ldots E_n$ then

$$P(E_{k/E}) = \frac{P(E_k)P(E/E_k)}{\sum_{i=1}^{n} P(E_i)P\left(\frac{E}{E_i}\right)}$$

Where k = 1, 2, ...n

2) **Mutually Exclusive and Exhaustive Events:** There are two mutually exclusive and exhaustive events A_1 and A_2 and the occurrence of the event B is in combination with A_1 and A_2, then for event A the conditional probability, given the event B is as under,

$$P(A_1/B) = \frac{P(A_1)P(B/A_1)}{P(A_1)P(B/A_1) + P(A_2)P(B/A_2)};$$

$$P(A_2/B) = \frac{P(A_2)P(B/A_2)}{P(A_1)P(B/A_1) + P(A_2)P(B/A_2)}$$

Example 10: Hypertension: By an automated blood-pressure machine, assume 84% of hypertensives and 23% of normotensives are categorised as hypertensive. What is the PV^+ and PV^- of the machine, assuming 20% of the adult population is hypertensive?

Solution: Sensitivity=0.84 and Specificity = 1–0.23= 0.77. Therefore, from the Bayes' rule:

$$PV^+ = P(B|A) = \frac{P(A|B) \times P(B)}{P(A|B) \times P(B) + P(A|\bar{B}) \times P(\bar{B})}$$

In other words, this can also be written as:

$$PV^+ = \frac{\text{Sensitivity} \times x}{\text{Sensitivity} \times x + (1 - \text{Specificity}) \times (1 - x)}$$

Scanned with CamScanner

Where x = P(B) = Prevalence of disease in reference population.

$$PV^{-} = \frac{\text{Specificity} \times (1-x)}{\text{Specificity} + (1-x) \times (1-\text{Sensitivity}) \times x}$$

Predictive value Positive (PV^{+}) = (.84) (0.2)/[(0.84) (0.2) + (0.23)(0.8)] =0.168/0.352 = 0.48

Also, Predictive value Positive PV^{-} = (0.77)(0.8)/[(0.77)(0.8)+(0.16)(0.2)]
=0.616/0.648 = 0.95

Example 11: A bag A contains 3 white and 4 black balls and a bag B contains 5 white and 3 black balls. One ball is drawn at random from one of the bags and it is found to be black. Find the probability that it was drawn from the bag B.

Solution: Let, X_1 = The event of ball being drawn from bag A,
X_2 = The event of ball being drawn from bag B and
X = The event of ball being black.

Since both the bag are equally likely to be selected, therefore

$$P(X_1) = P(X_2) = \frac{1}{2}$$

and $P(X/X_1) = 4/7$ and $P(X/X_2) = 3/8$

∴ Required Probability

$$P(X_2/X) = \frac{P(X_2).P(X/X_2)}{P(X_1).P(X/X_1) + P(X_2).P(X/X_2)}$$

$$= \frac{\frac{1}{2}\times\frac{3}{8}}{\frac{1}{2}\times\frac{4}{7}+\frac{1}{2}\times\frac{3}{8}} = \frac{\frac{3}{16}}{\frac{2}{7}+\frac{3}{16}} = \frac{\frac{3}{16}}{\left(\frac{32+21}{112}\right)} = \frac{3}{16}\times\frac{112}{53} = \frac{21}{53} = 0.396$$

Example 12: In a recruitment test, a student either guesses or copies or knows the answer to a multiple choice question with four choices. The probability that he makes a guess is $\frac{1}{4}$ and the probability that he copies the answer is $\frac{1}{5}$. The probability that his answer is correct, given that he copied it is $\frac{1}{7}$. Find the probability that he knew the answer to the question, given that he correctly answered it.

Solution: Let X_1 be the event that the answer is guessed, X_2 be the event that the answer is copied, X_3 be the event that the examinee knows the answer and X be the event that the examinee answers correctly.

Given, $P(X_1) = \frac{1}{4}, P(X_2) = \frac{1}{5}$,

Assume that events X_1, X_2 and X_3 are exhaustive.

∴ $P(X_1) + P(X_2) + P(X_3) = 1$

⇒ $P(X_3) = 1 - P(X_1) - P(X_2)$

$$P(X_3) = = 1 - \frac{1}{4} - \frac{1}{5} = \frac{20-5-4}{20} = \frac{11}{20}$$

Now $P(X/X_1)$ = Probability of getting correct answer by guessing = $\frac{1}{4}$ (Since 4 alternatives)

$P(X/X_2)$ = Probability of answering correctly by copying = $\frac{1}{7}$

And $P(X/X_3)$ = Probability of answering correctly by knowing = 1

Clearly, (X_3/X) is the event that he knew the answer to the question given that he correctly answered it.

$$\therefore P(X_3/X) = \frac{P(X_3).P(X/X_3)}{P(X_1).P(X/X_1)+P(X_2).P(X/X_2)+P(X_3).P(X/X_3)}$$

$$= \frac{\frac{11}{20}\times 1}{\frac{1}{4}\times\frac{1}{4}+\frac{1}{5}\times\frac{1}{7}+\frac{11}{20}\times 1} = \frac{\left(\frac{11}{20}\right)}{\left(\frac{1}{16}+\frac{1}{35}+\frac{11}{20}\right)} = \frac{\left(\frac{11}{20}\right)}{\left(\frac{35+16+308}{560}\right)} = \frac{\left(\frac{11}{20}\right)}{\left(\frac{359}{560}\right)}$$

$$= \frac{11}{20}\times\frac{560}{359} = \frac{11\times 28}{359} = \frac{308}{359} = 0.857$$

Example 13: A company produces plastic bottles and it is found that in production process 20% items are defective. Each item is inspected before being supplied to customers but the inspector incorrectly classifies an item 20% of the time. Only items classified as good are supplied. 816 items in all are supplied, how many of them are expected to be defective.

Solution: There are two cases:

1) The item is good and the inspector also declares it good for which the Probability

$$= \frac{8}{10}\times\frac{8}{10} = \frac{64}{100} \quad \text{.....(1)}$$

2) The item is defective and the inspector declares it good for which the Probability

$$= \frac{2}{10}\times\frac{2}{10} = \frac{1}{25} \quad \text{.....(2)}$$

The number of defective items supplied in a lot of 816 items

$$= 816\times\frac{\left(\frac{1}{25}\right)}{\left(\frac{64}{100}+\frac{1}{25}\right)} = 816\times\frac{\left(\frac{1}{25}\right)}{\left(\frac{16}{25}+\frac{1}{25}\right)} = 816\times\frac{\left(\frac{1}{25}\right)}{\left(\frac{17}{25}\right)} = 816\times\left(\frac{1}{25}\times\frac{25}{17}\right) = 48$$

6.1.8. Problems

Example 14: One coupon is drawn at random from a bag containing 26 coupons numbered 1 to 26. Represent the sample space and the event of drawing a coupon containing number which is a prime. Also find the number of elements in them.

Solution: Let S be the sample space and E be the event of occurrence a prime number.

Then S = {1, 2, 3, 4, 5,.... 26}

and E = {2, 3, 5, 7, 11, 13, 17, 19, 23}

also n (S) = 26 and n (E) = 9

Scanned with CamScanner

Example 15: Find the probability that a card selected at random from a deck is an ace or a queen.

Solution: The appearance of an ace card, E_1, and a queen card, E_2, represents mutually exclusive events since a given card can be only one of these types and not both. Accordingly,

$$P(E_1 \cup E_2) = P(E + E_1) + P(E_2)$$
$$= \frac{4}{52} + \frac{4}{52} = \frac{8}{52} = \frac{2}{13}$$

Example 16: A bag contains 8 red, 10 blue and 12 black balls. If one ball is drawn at random, find the probability that it is black.

Solution: Number of Red Balls = 8
Number of Blue Balls = 10
Number of Black Balls = 12
∴ Total Number of Balls = 8 + 10 + 12 = 30

The number of drawing 1 balls out of 30 = $^{30}C_1 = \frac{30!}{29! \times 1!} = 30$

∴ n (S) = 30

The number of ways of drawing 1 black balls out of 12
$= {}^{12}C_1 = 12$

∴ The Required Probability = $\frac{n(E)}{n(S)} = \frac{12}{30} = 0.4$

Example 17: In an examination hall there are 20 boys and 10 girls. Three students are selected at a random. What is the probability that 1 boy and 2 girls are selected?

Solution: No. of ways to select 1 boy from 20 boys = $^{20}C_1$

No. of ways to select 2 girls from 10 girls = $^{10}C_2$

No. of ways to select 3 students from 30 students = $^{30}C_3$

Probability of selecting 1 boy and 2 girls

$$= \frac{{}^{20}C_1 \times {}^{10}C_2}{{}^{30}C_3} = \frac{\frac{20!}{19!} \times \frac{10!}{8!2!}}{\frac{30!}{27! \times 3!}} = \frac{20 \times 45}{5 \times 29 \times 28} = \frac{45}{203}$$

Example 18: Two cards are drawn from a standard pack of cards at random. What is the probability that it will be:
i) A Diamond and a Heart, ii) iA King and a Queen, iii) Two Kings

Solution: From a pack of 52 cards, two cards can be drawn in $^{52}C_2$ ways, i.e., the exhaustive number of cases = $^{52}C_2 = \frac{52 \times 51}{2 \times 1} = 1326$.

i) Favourable number of cases for drawing a diamond and a heart = $^{13}C_1 \times {}^{13}C_1 = 169$
Required Probability
$$= \frac{\text{Favourable No. of Cases}}{\text{Exhaustive No. of Cases}} = \frac{{}^{13}C_1 \times {}^{13}C_1}{{}^{52}C_2} = \frac{169}{1326}$$

ii) The favourable number of cases for drawing a king and a queen
$= {}^4C_1 \times {}^4C_1 = 16$

Required Probability = $\frac{{}^4C_1 \times {}^4C_1}{{}^{52}C_2} = \frac{16}{1326} = \frac{8}{663}$

iii) 2 kings can be drawn out of 4 kings in $^4C_2$ ways, i.e., the favourable number of cases
$= {}^4C_2 = \frac{4 \times 3}{2 \times 1} = 6$

Therefore, Required Probability = $\frac{6}{1326} = \frac{1}{221}$

6.2. PROBABILITY DISTRIBUTION

6.2.1. Introduction

A probability distribution is a statistical model that shows the possible outcomes of a particular event or course of action as well as the statistical likelihood of each event. **A probability distribution** for a particular random variable is a function or table of values that links each outcome in the sample space of a statistical experiment with its probability of occurrence. A theoretical probability distribution is a theoretical model of the relative frequencies of events in a population. Suppose an unbiased coin is flipped twice. This simplest statistical experiment can have four distinct outcomes, namely HH, HT, TH and TT. If X represents the number of heads that result from the experiment, then the table 6.1 sufficiently represents the probability distribution, X being a random variable and taking values 0, 1 and 2.

Table 6.1: Probability Distribution of the Random Variable (X)

Number of Heads (X)	Probability
0	0.25
1	0.50
2	0.25

The range of feasible values that a random variable can attain and the probability that the value of the random variable is within any (measurable) subset of that range is explained by the probability distribution. When the values are taken by the random variable, which are the subset of real numbers, then the cumulative distribution function completely describes the probability distribution. The value at each real X is the probability that the random variable is smaller than or equal to X. The probability for each value of the random variable is given by a probability distribution in the form of a graph, table, or formula. If X is a random variable then the probability that X occurs is denoted by P(X). It must be the case that $0 \le P(X) \le 1$ for each value of X and $\sum P(X) = 1$ (the sum of all the probabilities is 1).

6.2.2. Applications of Probability Distribution

1) The use of probability distribution is to calculate confidence intervals for parameters and to calculate critical regions for "test of hypothesis".
2) It is useful to determine a reasonable distributional level/model for univariate data.
3) Statistical intervals and "tests of hypothesis" are often based on specific distribution assumptions. Before computing an interval or test based on a distributional assumption, we need to verify that the assumption is justified for the given data set. Here, the distribution does not need to be the best-fitting distribution for the data, but a sufficient model to yield valid conclusions for the statistical technique.
4) Simulation studies with random numbers generated from using a specific probability distribution are often useful.

Scanned with CamScanner

6.2.3. Types of Probability Distribution

The various types of Probability Distributions are as shown in **figure below**:

1) **Discrete Probability Distribution:** In a probability distribution of a random variable X, in which X can only take the values of discrete integers, then it is called discrete probability distribution. A probability distribution is called discrete probability distribution when its random variable is a discrete variable. The following examples illustrate discrete probability distributions:
   i) A car can have only 0, 1, 2, 3 or 4 flat tyres.
   ii) A bookshop has only 0, 1, 2, 3, 4 or 5 copies of a particular title in stock.
   iii) The number of employees absent on a given day is 0, 1, 2, 3, etc.

Types of Discrete Probability Distribution

i) Binomial Distribution, and
ii) Poisson Distribution.

2) **Continuous Probability Distribution:** The probability distribution of a random variable is called **Continuous Probability Distribution** if the given random variable is continuous. Continuous variables contain very large number of outcomes (which are in fact uncountable) but discrete variables contain finite number of outcome values. In continuous random variables, the probability of a single outcome will be close to zero since the total probability '1' is divided among a very large number of outcomes. For continuous random variables, calculation of probability for a single value may not be significant.

Some examples that will clarify the difference between discrete and continuous variables are:

1) Suppose that the Indian air force sets the qualification that all pilots must weigh between 55kg and 65kg. The weight of a pilot would be an example of a continuous variable; since a pilot's weight could take on any value between 55kg and 65kg.

2) Suppose one flips a coin and counts the number of tails. The number of tails could be any integer value between 0 and plus infinity. One could not, e.g., get 2.5 tails. Therefore, the number of tails must be a discrete variable.

Types of Continuous Probability Distribution

The following continuous probability distributions are illustrated:

1) Uniform Distribution
2) Exponential Distribution
3) Normal Distribution
4) Student's t Distribution
5) Chi-Square Distribution
6) F Distribution

6.2.4. Binomial Distribution

A binomial distribution takes place when there are only "two" mutually exclusive possible outcomes. **For example,** the flipping of a coin may only result in either a head or a tail. If we refer to one outcome as 'success' then we have to term the other outcome as 'failure'. This binomial distribution is also referred to as Bernoulli distribution after the Swiss Mathematician **Jacob Bernoulli** who derived it.

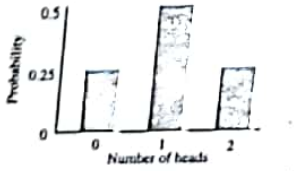


Figure 6.3: Probabilities of 0, 1 and 2 Heads

The **figure 6.3** shows the example of coins.

Table: Probabilities of Getting 0, 1 or 2 Heads

Number of Heads	Probability
0	0.25
1	0.5
2	0.25

Definition: *If the probability of a successful trial is p and that of failure in any trial is q (= 1 – p), then the probability of having x successful outcomes in an experiment of n independent trials is as follows:* $f(x) = {}^nC_x p^x q^{n-x}$

This probability distribution is called the ***binomial probability distribution.*** *The binominal distribution is a discrete distribution with parameters n and p. If p and q are equal it is symmetrical, otherwise it is nonsymmetrical.*

Notation for Binomial Probability

1) **r:** The number of successes that result from the binomial experiment.
2) **n:** The number of trials in the binomial experiment.
3) **p:** The probability of success of an individual trial.
4) **q:** The probability of failure of an individual trial. (This is equal to 1 – p.)
5) **b(x; n, p):** Binomial probability – the probability that n-trial binomial experiment results in exactly x successes, when the probability of success on an individual trial is p.
6) ^nC_r: The number of combinations of n things, taken r at a time.

Probability Table for a Binomial Distribution

Number of Success	Probability Distribution	Number of Success	Probability Distribution
0	$^nC_0 p^0 q^{n-0} = q^n$	n	$^nC_n p^n q^{n-n}$
1	$^nC_1 p^1 q^{n-1}$	n–1	$^nC_{n-1} p^{n-1} q$
2	$^nC_2 p^2 q^{n-2}$	n–2	$^nC_{n-2} p^{n-2} q^2$
3	$^nC_3 p^3 q^{n-3}$	n–3	$^nC_{n-3} p^{n-3} q^3$

6.2.4.1. Assumptions of Binomial Distribution

1) The number of observations is fixed.
2) Each trial has two mutually exclusive possible outcomes, i.e., success or failure.
3) Each trial is independent of other trials.
4) Each trial has the constant probability of success (say p) or failure (= 1 – p).

6.2.4.2. Properties of Binomial Distribution

1) A discrete probability distribution which is based on Binomial Theorem is called Binomial Distribution.
2) Total number of trials 'n' and probability of success in each trial 'p' are the parameters of the distribution.
3) The mean and variance of a binomial distribution are 'np' and 'npq' respectively.
4) Mean is always greater than its variance for binomial distribution.
5) The binomial distribution is called symmetric distribution when p=1/2.
6) The binomial distribution is called negatively skewed distribution if p>1/2 & if p<1/2, then it is a positively skewed distribution.
7) Binomial Distribution tends to Poisson Distribution if 'n' is very large and 'p' is very small.
8) Binomial Distribution tends to Normal Distribution if 'n' is very large; p and q are not small.

Scanned with CamScanner

6.2.4.3. Applications of Binomial Distribution

1) It is used to find probability of getting 'x' success in 'n' independent Bernoulli trials.
2) It is also used in sampling, inspection plans, genetic experiments, for finding control limits for p and 'np' charts in statistical quality control.

6.2.4.4. Fitting a Binomial Distribution

When a binomial distribution has to be fitted to an observed data, following is the procedure:

1) Determine mean of the given distribution and then find the values of p and q. If one of these values is known, the other can be found out by the simple relationship $p + q = 1$.
2) Expand the binomial $(q + p)^n$. The number of terms in the expanded binomial is $(n + 1)$.
3) Multiply each term of the expanded binomial by N (total frequency), to get the expected frequency in each category.

6.2.4.5. Problems

Example 19: Fit a binomial distribution to the following distribution of 156 samples:

No. of Defective Items	7	6	5	4	3	2	1	0
No. of Samples	1	6	32	36	48	24	7	2

Solution: To fit the binomial distribution to the data, we need the values of n, N, p and q.

X	f	fX
0	2	0
1	7	7
2	24	48
3	48	144
4	36	144
5	32	160
6	6	36
7	1	7
	$\sum f = 156$	$\sum fX = 546$

Here, $n = 7$, $N = 156$

$$\text{Mean} = np = \bar{X} = \frac{\sum fX}{\sum f} = \frac{546}{156} = 3.5$$

$$\Rightarrow \quad 7 \times p = 3.5$$

$$\Rightarrow \quad p = \frac{1}{2}$$

$$\therefore \quad q = 1 - \frac{1}{2} = \frac{1}{2}$$

Computations of Frequencies

No. of defective	$f(x) = {}^nC_x p^x q^{n-x}$	Expected frequency = N × f(x)
0	$P(0) = {}^7C_0\left(\frac{1}{2}\right)^0\left(\frac{1}{2}\right)^7 = \frac{1}{128}$	1.22
1	$P(1) = {}^7C_1\left(\frac{1}{2}\right)^1\left(\frac{1}{2}\right)^6 = \frac{7}{128}$	8.53
2	$P(2) = {}^7C_2\left(\frac{1}{2}\right)^2\left(\frac{1}{2}\right)^5 = \frac{21}{128}$	25.6
3	$P(3) = {}^7C_3\left(\frac{1}{2}\right)^3\left(\frac{1}{2}\right)^4 = \frac{35}{128}$	42.65
4	$P(4) = {}^7C_4\left(\frac{1}{2}\right)^4\left(\frac{1}{2}\right)^3 = \frac{35}{128}$	42.65
5	$P(5) = {}^7C_5\left(\frac{1}{2}\right)^5\left(\frac{1}{2}\right)^2 = \frac{21}{128}$	25.6
6	$P(6) = {}^7C_6\left(\frac{1}{2}\right)^6\left(\frac{1}{2}\right)^1 = \frac{7}{128}$	8.53
7	$P(7) = {}^7C_7\left(\frac{1}{2}\right)^7\left(\frac{1}{2}\right)^0 = \frac{1}{128}$	1.22

Example 20: Infectious Disease: The most common and routine laboratory tests done on medical examination is blood count. There are two main features of blood count are as given below:

1) Calculating the number of white blood cells ("white count") and
2) Differentiating white blood cells which do exist into 5 classes – neutrophils, lymphocytes, monocytes, eosinophils, and basophils. It is known as "differential".

White count and differential are used together widely in clinical diagnoses. We generally focus on the differential, i.e., the distribution of the number of neutrophils k out of 100 white blood cells.

i) What is the probability that 2nd and 5th cells taken will be neutrophils and the remaining cells non-neutrophils? It is given a probability of 0.6 for any one cell is a neutrophil.
ii) Find out the probability that any two cells out of five will be neutrophils.

Solution:

i) Suppose 'x' is the notation of neutrophil and 'o' is the notation of non-neutrophil. Since probabilities of success and failure are given by 0.6 and 0.4 respectively, and different cells are supposed to be independent. Hence, the probability is:
$$q \times p \times q \times q \times p = p^2q^3 = (0.6)^2(0.4)^3$$

ii) There is only one 'oxcox' out of the 10 possible orderings that results in two neutrophils. Ten possible orderings are given in table 6.2.

Table 6.2: Probable Orderings for 2 Neutrophils of 5 Cells

xxooo	oxxoo	ooxox
xoxoo	oxoxo	oooxx
xooxo	oxoox	
xooox	ooxxo	

The number of ordering= The number of ways to select 2 cells to be neutrophils out of 5 cells= ${}^5C_2 = (5\times4)/(2\times1) = 10$.

Probability of some of orderings in **Table 6.2** is same as that for the ordering 'oxoox', that is, $(0.6)^2(0.4)^3$. Hence, probability of getting two neutrophils in 5 cells is as follows:
$= {}^5C_2 = (5\times4)/(2\times1) = 10\ (0.6)^2(0.4)^3$
$= 0.230$

Example 21: Let us consider that a bomb dropped from an aeroplane hits an enemy's bridge then its probability is 0.4. Two bombs cutting the bridge can destroy. In case six bombs are dropped from the aeroplane then determine the probability that the bridge is destroyed.

Solution: Given that, the probability of hitting the target is 0.4.
Assume that P(S) is the notation of it, then:

$$P(S) = 0.4 = \frac{2}{5}.$$

The probability of not hitting the target is $P(F) = \frac{3}{5}$

We have, n = 6, suppose that the X is a random variable target will destroyed in case, 2 bombs hit the target.

∴ Probability that the target will be destroyed is given as:
$P(X \geq 2) = 1 - [P(X = 0) + P(X = 1)]$

Scanned with CamScanner

$$= 1 - \left\{ \left(\frac{3}{5}\right)^6 + {}^6C_1 \left(\frac{2}{5}\right)^1 \left(\frac{3}{5}\right)^5 \right\} = 1 - \left\{ \left(\frac{3\times3\times3\times3\times3\times3}{5\times5\times5\times5\times5\times5}\right) + 6\times\frac{2}{5}\times\left(\frac{3\times3\times3\times3\times3}{5\times5\times5\times5\times5}\right) \right\}$$

$$= 1 - \left\{ \frac{729}{15625} + \frac{2912}{15625} \right\} = 1 - \left\{ \frac{3641}{15625} \right\} = \frac{15625 - 3641}{15625} = \frac{11984}{15625} = 0.767$$

Example 22: If there is one vessel on an average in every ten is wrecked, determine the probability that out of five vessel likely to reach, minimum 4 will arrive safely.

Solution: Given that the probability of wrecking a vessel is $\frac{1}{10}$.

$\therefore\ P = \frac{1}{10}, q = \frac{9}{10}, n = 5.$

Or minimum 4 arrive safely this means 1 is wrecked or two wrecked or three wrecked.

$\therefore$ Required probability = P(0) + P(1) + P(2) + P(3)

$$= {}^5C_0 q^5 + {}^5C_1 q^4 p + {}^5C_2 q^3 p^2 + {}^5C_3 q^2 p^3$$

$$= \left(\frac{9}{10}\right)^5 + 5\left(\frac{9}{10}\right)^4\left(\frac{1}{10}\right) + \frac{5.4}{2.1}\left(\frac{9}{10}\right)^3\left(\frac{1}{10}\right)^2 + \frac{5.4}{2.1}\left(\frac{9}{10}\right)^2\left(\frac{1}{10}\right)^3$$

$$= \frac{9^2}{10^5}(729 + 405 + 90 + 10) = \frac{81(1134)}{10^5} = 0.91854$$

6.2.5. Poisson Distribution

Poisson distribution was developed by the French mathematician **Simeon Denis Poisson** in 1837 (1781–1840). Poisson distribution arises when you count a number of events across time or over an area. Poisson distribution is a discrete random variable distribution that expresses probability of the given number of events in a fixed interval of time, if these events occur with a known average rate and independent of the time since the last event. Poisson distribution can also be used for the number of events in other specified intervals such as distance, area or volume. Poisson distribution can be used as an alternative to Binomial distribution in the case of very large samples. In binomial distribution the value of n is finite. But there are few problems when it is not possible to estimate 'n'. In this case binomial distribution is inapplicable because one cannot expand $(p + q)^n$, if n is not known. In such cases Poisson distribution is applicable.

Examples of Poisson Distribution

1) The number of telephone calls received at a particular switch board per minute during a certain hour of the day.
2) The number of accidents per year in a district of Maharashtra.
3) The number of trucks passing a certain point per minute.
4) The number of persons born deaf and dumb per year in a city.
5) The number of typing errors per page.
6) The number of printing errors per page in a book.
7) The number of defective blades in a pack of 1000.

6.2.5.1. Definition and Probability Function

It is defined by the probability function: $f(x) = \frac{e^{-m} m^x}{x!}$

for x (no. of successes) = 0, 1, 2, 3, ...where m is a fixed positive number.

e = 2.71828 (app.)

The successive terms of the distribution for different values of x are as follows:

x:	0	1	2	3	Total
f(x)	e^{-m}	$e^{-m} \times m$	$\frac{e^{-m} \times m^2}{2!}$	$\frac{e^{-m} \times m^3}{3!}$	1 (as the total probability must be unity)

Notation for Poisson Probability

The following notation is helpful, when we talk about the Poisson distribution.

1) **e:** A constant equal to approximately 2.71828. (Actually, e is the base of the natural logarithm system.)
2) **m:** The mean number of successes that occur in a specified region.
3) **x:** The actual number of successes that occur in a specified region.
4) **P(x; m):** The Poisson probability that exactly x successes occur in a Poisson experiment, when the mean number of successes is μ.

6.2.5.2. Assumptions of Poisson Distribution

1) The number of events is discrete in a given interval, say [0, t].
2) The probability is approximately proportional to the length of the interval for an event which may occur in a short interval $[t, t + \Delta t]$.
3) The occurrences of events are independent in non-overlapping intervals.
4) The probability of two events is negligible in a short interval $[t, t + \Delta t]$.

6.2.5.3. Properties of Poisson Distribution

1) The outcomes that occur in the result of the experiment can be classified as successes or failures.
2) The average number of successes (μ) which occur in a specified region is known.
3) The probability that a success will occur is directly related to the size of the region.
4) The probability that a success will occur in an extremely small region is almost zero.
5) It is a discrete probability distribution where the random variable x assumes the infinite set of values 0, 1, 2....
6) i) Mean (m) = parameter of the distribution. ii) Variance $(\sigma^2) = m$ iii) Standard Deviation $(\sigma) = \sqrt{m}$

   iv) Skewness = $\frac{1}{\sqrt{m}}$ v) Kurtosis (Peakedness) = $\frac{1}{m}$
7) The mode of Poisson distribution is that value of x which occurs with largest probability. It may have either one or two modes. If 'm' is not an integer, the mode in the integral value between (m – 1) and (m). If, however, 'm' is an integer, then there are two modes which are (m – 1) and (m).
8) If x and y are two independent Poisson variates with parameters m_1 and m_2 respectively, then their sum (x + y) is also a Poisson variate with parameter $m_1 + m_2$.
9) The first, second, and third row moments are respectively (m), $(m^2 + m)$, $(m^3 + 3m^2 + m)$.

6.2.5.4. Applications of Poisson Distribution

1) To count the number of defects of an item used in quality control statistics.
2) To count the number of germs in a substance.
3) To count the number of particles discharged from a radioactive substance in a nuclear reactor.
4) To count the number of casualties with the customer of an insurance company.
5) In queuing problems to count the number of arrivals of customers at a billing counter.
6) Number of traffic arrivals such as aeroplanes at airports, ships at harbours, trucks at the gates, etc.

Scanned with CamScanner

7) In determining the total number of deaths carried by a specific disease (and not epidemic), in a particular locality or a particular (specified) period of time.
8) The number of typing errors per page in a typed material.
9) In problems dealing with the inspection of manufactured products with the probability that any one piece is defective is very small and the lots are very large.
10) To model the distribution of the number of persons joining a queue to receive a service or purchase of a product.

6.2.5.5. Fitting a Poisson Distribution

The process of fitting a Poisson distribution is very simple. First one has to find the value of m, i.e., the average occurrence, and calculate the frequency of 0 successes. The other frequencies can be calculated as follows:

$$f(x)=\frac{e^{-m}m^x}{x!} \text{ and } f(x+1)=\frac{e^{-m}m^{(x+1)}}{(x+1)!}$$

$$\Rightarrow \quad \frac{f(x+1)}{f(x)}=\frac{m^{(x+1)}}{(x+1)!}\times\frac{x!}{m^x}=\frac{m}{(x+1)}$$

$$\Rightarrow \quad f(x+1)=\frac{m}{(x+1)}f(x)$$

When, $x=0 \Rightarrow f(1)=m\times f(0)$; When, $x=1 \Rightarrow f(2)=\frac{m^2}{2}f(1)$; When, $x=2$

$\Rightarrow f(3)=\frac{m^3}{3}f(2)$; and so on.

This recurring relation provides a very convenient method for fitting the Poisson distribution. The only probability we need to know is $f(0)=e^{-m}$, where m is the only parameter of the Poisson distribution. By multiplying N (the total frequency) each probability of the Poisson distribution, we get the expected frequencies for respective probabilities.

6.2.5.6. Problems

Example 23: *If 4% of the electric geysers manufactured by a company are defective, use Poisson distribution to find the probability in a sample of 100 geysers when:*

i) None is defective *ii) 5 geysers are defective*

(Given: $e^{-4}=0.018$)

Solution: *Let p = probability of a defective geyser* $=4\%=\frac{4}{100}=0.04$

We are given n = 100, p = 0.04; $\therefore$ *Mean (m) = np = 100 × 0.04 = 4*

The Poisson distribution is given as:

$$P(X=x)=\frac{e^{-m}\cdot m^x}{x!}$$

1) **None are defective**

$$P(X=0)=\frac{e^{-4}\times 4^0}{0!}=e^{-4}=0.018$$

$$P(0)=0.018$$

2) **5 geysers are defective**

$$P(X=5)=\frac{e^{-4}\times 4^5}{5!}=\frac{0.018\times 1024}{5\times 4\times 3\times 2}=\frac{18.432}{120}=0.154$$

$$P(5)=0.154$$

Example 24: Infectious Disease: Understand through the example of typhoid-fever. Assume that the over one year period, the number of deaths from typhoid fever is Poisson distributed with the parameter $\mu = 4.6$. Find the probability distribution of the number of deaths over a 6-month period and a 3-month period?

Solution:

1) Suppose X is the number of death due to typhoid fever in six months. Since $\mu = 4.6$, t = 1 year, it follows that $\lambda = 4.6$ death/year. For a six month period we have $\lambda = 4.6$ deaths per year, $t = 0.5$ year. Therefore, $\mu = \lambda t = 2.3$. So,

$$P(X=0)=e^{-2.3}=.100 \qquad P(X=1)=\frac{2.3}{1!}e^{-2.3}=.231$$

$$P(X=2)=\frac{2.3^2}{2!}e^{-2.3}=.265 \qquad P(X=3)=\frac{2.3^3}{3!}e^{-2.3}=.203$$

$$P(X=4)=\frac{2.3^4}{4!}e^{-2.3}=.117 \qquad P(X=5)=\frac{2.3^5}{5!}e^{-2.3}=.054$$

$$P(X\geq 6)=1-(.100+.231+.265+.203+.117+.054)=.030$$

2) Assume that Y is the number of deaths in three months. For a period of three month, $\lambda = 4.6$ deaths/ year, $t = .25$ year, $\mu = \lambda t = 1.15$. So,

$$P(Y=0)=e^{-1.15}=.317 \qquad P(Y=1)=\frac{1.15}{1!}e^{-1.15}=.364$$

$$P(Y=2)=\frac{1.15^2}{2!}e^{-1.15}=.209 \qquad P(Y=3)=\frac{1.15^3}{3!}e^{-1.15}=.080$$

$$P(Y\geq 4)=1-(.317+.364+.209+.080)=.030$$

In figure 6.4, distributions are plotted.

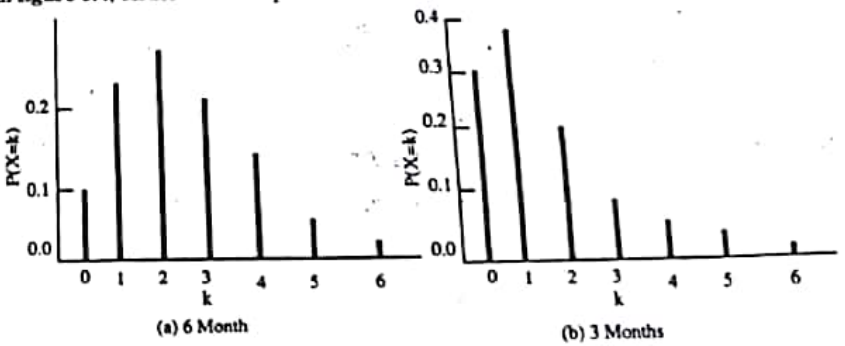


Figure 6.4: Distribution of the Number of Deaths due to Typhoid Fever over various Time Intervals

Example 25: Bacteriology supposing A = 100 cm^2 and $\lambda = .02$ colonies/cm^2, find the probability distribution of the number of bacterial colonies.

Scanned with CamScanner

Solution: Given $\mu = \lambda A = 100(.02) = 2$.
Assume X is the number of colonies, then:

$P(X=0) = e^{-2} = 0.135$

$P(X=1) = e^{-2}\, 2^1/1! = 2e^{-2} = 0.271$

$P(X=2) = e^{-2}\, 2^2/2! = 2e^{-2} = 0.271$

$P(X=3) = e^{-2}\, 2^3/3! = \frac{4}{3}e^{-2} = 0.180$

$P(X=4) = e^{-2}\, 2^4/4! = \frac{2}{3}e^{-2} = 0.090$

$P(X \geq 5) = 1 - (0.135+0.271+0.271+0.180+0.90) = 0.053$

Obviously, the larger λ is, the more bacterial colonies we would probable to catch.

6.2.6. Normal Distribution

A family of continuous probability distributions described by the normal equation is called the normal distribution.

Normal Equation
The normal distribution is defined by the following equation:

The value of the random variable y is:

$$y = f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{1}{2\sigma^2}(x-\mu)^2}$$

Where, x = Normal random variable, μ = Mean, σ = Standard deviation, $\pi = 3.14159$, and $e = 2.71828$.

The random variable x in the normal equation is called the normal random variable. The probability density function for the normal distribution is called the normal equation. The graph of normal equation is called the **normal curve** which is a bell-shaped curve that extends indefinitely in both directions, coming closer and closer to the horizontal axis without even touching it.

6.2.6.1. Definition and Probability Function

A continuous random variable x is said to have a normal distribution with parameters μ (mean) and σ^2 (variance), if it has the probability density function represented by the equation:

$$y = f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{1}{2\sigma^2}(x-\mu)^2} \quad -\infty < x < +\infty$$

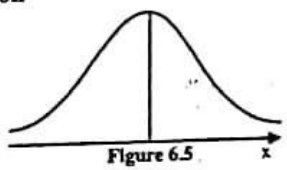


Figure 6.5

Where,
y = The height of the curve at a distance x from mean; it indicates the probability density of that particular value of the variable x,
σ = Standard deviation of the given normal distribution,
μ = Mean of the random variable x,
π = The constant = 3.1416; $\sqrt{2\pi} = 2.5066$, and
e = The constant = 2.1783 (base of natural logarithm).

Probability density can be converted into frequency density by multiplying the probability density equation by N (i.e. the total number of items in the distribution). The curve which is representing the normal distribution in the **figure 6.5** is known as normal curve. The probability that x lies between the ordinates x = a, x = b, (b > a), i.e., probability (a < x < b) is represented by the area under the curve between a and b. The total area covered by the x-axis and the curve is one.

The diagram of f(x) is a famous **'bell shaped'** curve. The top of the curve is at the top of the point where the mean lies. The two tails of the curve on the right and left of the mean (μ) extend to infinity which means that they do not intersects x-axis. The curve tends to flatten out in the condition when the value of standard deviation is large. On the other hand, it has a sharp peak when the value of standard deviation is small.

6.2.6.2. Properties of Standard Normal Distribution

1) It is a continuous probability distribution having parameters μ and σ.
2) The normal curve is perfectly symmetrical about the mean (μ) and is bell shaped. The two tails of the curve on either side of mean extend to infinity.
3) Mean = Median = Mode, Skewness = 0.
4) It has only one mode.
5) Q_1 and Q_3 are equidistant from median (or mean)
   $Q_1 = m - 0.67 \times \sigma$, $Q_3 = m + 0.67 \times \sigma$ (app.)
   Quartile Deviation = $0.67 \times \sigma$, mean deviation (about mean) = $4/5 \times \sigma$.
6) The maximum ordinate is at x = m, and its value is $\frac{1}{\sigma\sqrt{2\pi}}$. Thus, probability is inversely proportional to Standard Deviation.
7) The area under the curve means the area lying between the curve and the horizontal axis and is equal to the number of frequencies. The s.d. (σ) distributes this area as follows:
   i) 68.27% of the area lies between $x = -\sigma$ to $x = +\sigma$, 34.134% area will lie on either side of the mean.
   ii) 95.45% of the area lies between $x = -2\sigma$ to $x = +2\sigma$ i.e. 47.725% area will lie on either side of the mean.
   iii) 99.73% of the area lies between $x = -3\sigma$ to $x = +3\sigma$, 49.865% area will lie on either side of the mean.
8) x can lie anywhere between: $-\infty < x < +\infty$.

6.2.6.3. Standard Normal Distribution

A random variable that has a normal distribution with a mean of 0 and a standard deviation of 1 is said to have a standard normal distribution. We use the letter z to represent this particular normal random variable. Normal distributions can be transformed to standard normal distributions by the formula:

$$z = \frac{x - \mu}{\sigma}$$

Where,
x = Value from the original normal distribution, μ = Mean of the original normal distribution, and
σ = Standard deviation of original normal distribution.

The standard normal distribution is sometimes called the **z-distribution**.

Scanned with CamScanner

6.2.6.4. Normal Curve

The normal distribution graph is based on two factors namely, the mean and standard deviation. The mean of the distribution finds the position of the centre of the curve, and the standard deviation finds the height and width of the curve. When the value of standard deviation is high then the curve will be short in height and wide in width. When the value of standard deviation is small then the curve will be tall in height and narrow in width. All normal distributions look like a symmetric, bell-shaped curve, as shown in **figure 6.6**.

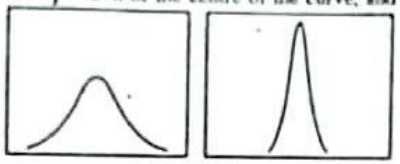


Figure 6.6

The curve on the left is shorter and wider than the curve on the right, because the curve on the left has a bigger standard deviation.

6.2.6.5. Probability and the Normal Curve

The normal distribution is a continuous probability distribution. These implications are as follows:

1) The total area under the curve is 1.
2) The probability that a normal random variable X equals any particular value is Zero.
3) The probability that X is greater than *a* equals the area under the normal curve bounded by 'a' (refer the figure above) and 'infinity plus' (as indicated by the non-shaded area in the **figure 6.7**).
4) The probability that X is less than 'a' equals the area under the normal curve bounded by 'a' and minus infinity (refer to the shaded area in the **figure 6.7**).

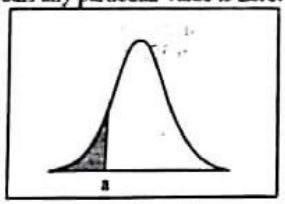


Figure 6.7: Normal Curve

Additionally, every normal curve (regardless of its mean or standard deviation) conforms to the following "rule":

1) About 68% of the area under the curve falls within 1 standard deviation of the mean.
2) About 95% of the area under the curve falls within 2 standard deviations of the mean.
3) About 99.7% of the area under the curve falls within 3 standard deviations of the mean.

6.2.6.6. Fitting a Normal Curve

For the fitting of a normal distribution to the given data, one has to first calculate the mean μ and standard deviation σ from the given data.

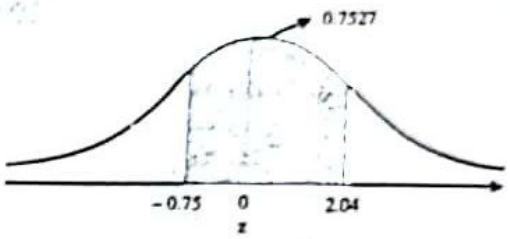
Then the normal curve fitted to the given data is given by:

$$y = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} \quad -\infty < x < +\infty$$

We first find the standard normal variates corresponding to the lower units of each class-interval to calculate the expected normal frequency. From the normal tables the areas under the normal curve are calculated for each standard normal variate. The areas for the successive class intervals are obtained by subtracting the successive areas and we get the normal frequencies by multiplying these areas with N.

6.2.6.7. Problems

Example 26: To find the area under standard normal curve between $z = -0.75$ and $z = 2.04$



Solution: Required area = (area between $z = -0.75$ and $z = 0$) + (area between $z = 0$ and $z = 2.04$)

= (area between $z = 0$ and $z = 0.75$) + (area between $z = 0$ and $z = 2.04$)

$= 0.2734 + 0.4793 = 0.7527$

Now $P(-0.75 \le z \le 2.04) = 0.7527$

Example 27: 2000 applicants appeared in an interview. Distribution of marks is assumed to be normal with mean (μ) = 30 and $\sigma = 6.25$. How many applicants are expected to get marks:

i) Between 20 and 40,
ii) Less than 35, and
iii) Above 50?

Solution: Given, $\mu = 30$, $\sigma = 6.25$

Since $z = \frac{x-\mu}{\sigma}$

i) Expectation of marks between 20 and 40.

Value of z when x = 20

$$z = \frac{20-30}{6.25} = -1.6$$

Value of z when x = 40

$$z = \frac{40-30}{6.25} = 1.6$$

Area between $z = 0$ and $z = -1.6$ is 0.4452

Area between $z = 0$ and $z = 1.6$ is 0.4452

$\therefore$ Area between $z = -1.6$ and $z = 1.6$ is $0.4452 + 0.4452 = 0.8904$

Required probability = 0.7108

Number of applicants with an expectation of marks between 20 and 40 = $2000 \times 0.7108 = 1421.6 \approx 1422$

ii) Expectation of marks less than 35.

$$\therefore \quad z = \frac{35-30}{6.25} = 0.8$$

$\therefore \quad P(z < 0.8) = 0.5 - 0.2881 = 0.2119$

Number of applicants with an expectation of marks less than 35 = $2000 \times 0.2119 = 423.8 \approx 424$

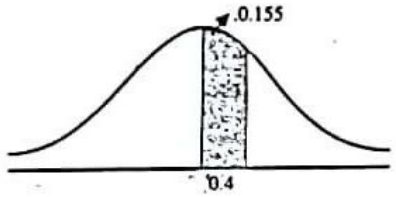
Scanned with CamScanner

iii) Expectation of marks above 50.

$$z=\frac{50-30}{6.25}=3.2$$

$P(z > 3.2) = 0.5 - 0.4993 = 0.0007$

Number of applicants with an expectation of marks above 50 = 2000 × 0.0007= 1.4 ≈ 1



Example 28: The daily income of 2,000 employees is normally distributed around a mean of ₹80 and with a standard deviation of ₹5. Estimate the number of employees whose daily income will be:

i) Between ₹ 80 and ₹82,
ii) More than ₹85,
iii) Less than ₹73,

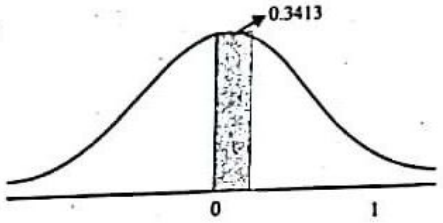
Solution:

i) When X = 80, $Z_1 = (80 - 80)/5 = 0$

When X = 82, $Z_2 = \frac{82-80}{5} = 0.4$

Area between Z = 0 and Z = 0.4 is 0.1554

Therefore the number of employees with daily income between ₹ 80 and ₹ 82 = 2000 × 0.1554 = 311

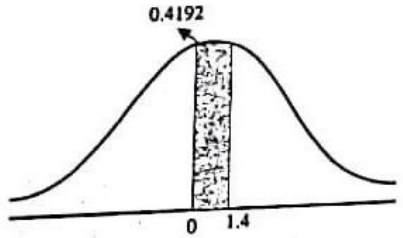


ii) When X = 85, $Z_2 = \frac{85-80}{5} = 1$

Area between Z = 0 and Z = 1 is 0.3413

Area greater than ₹85 = 0.5 − 0.3413 = 0.1587

Therefore, the number of employees with daily income more than ₹85 is 2,000 × 0.1587 = 317



iii) When X = 73, $Z_2 = \frac{73-80}{5} = -\frac{7}{5} = -1.4$

Area between Z = 0 and Z = 1.4 = 0.4192

Area less than ₹ 73 = 0.5 − 0.4192 = 0.0808

Therefore, the number of employees with daily income more than ₹ 73 is 2,000 × 0.0808 = 162

6.3. EXERCISE

6.3.1. Very Short Answer Type Questions

1) Define probability.
2) What is sample space?
3) Define random variable.
4) State the Baye's theorem.
5) Bring out the properties of probability.

6.3.2. Short Answer Type Questions

1) Define probability and bring out the importance of probability.
2) Define mutually exclusive events.
3) What is marginal probability?
4) Define conditional probability.
5) State the Bayes' theorem.

6.3.3. Long Answer Type Questions

1) In a bolt factory machines A, B and C manufacture respectively 25%, 35% and 40%. Of the total of their output 5, 4 and 2 per cent are defective bolts. A bolt is drawn at random from the product and is found to be defective. What are the probabilities that it was manufactured by machines A, B and C?
[Ans: A = 0.37; B = 0.40; C = 0.23]

2) Let A, B and C be three mutually and exhaustive events. Find P(B), if $\frac{1}{3}P(C)=\frac{1}{2}P(A)=P(B)$.
[Ans: $P(B) = \frac{2}{7}$]

3) If P(A) = 0.50, P(B) = 0.40 and $P(A \cup B) = 0.70$, find P(A/B) and $P(\overline{A} \cup B)$. State whether A and B are independent.
[Ans: $P(A/B) = \frac{1}{2}$, $P(\overline{A} \cup B) = 0.7$, independent]

4) Given $P(A) = \frac{1}{4}$, $P(A/B) = \frac{1}{4}$ and $P(B/A) = \frac{1}{2}$, find if
i) A and B are mutually exclusive.
ii) A and B are independent.
[Ans: (i) Not mutually exclusive, (ii) Independent]

5) Out of the numbers 1 to 100, one is selected at random. What is the probability that it is divisible by 7 or 8?
[Ans: $\frac{1}{4}$]

6) A committee of 4 people is to be appointed from 3 offices of the production department, 4 officers of the purchase department, two officers of the sales department and one chartered accountant.

Scanned with CamScanner

6) Find the probability of forming the committee in the following manner:
   i) There must be one from each category.
   ii) The committee should have atleast one from the purchase department.
   iii) The chartered accountant must be in the committee.

[Ans: (i) $\frac{8}{70}$, (ii) $\frac{13}{14}$, (iii) $\frac{2}{5}$]

7) The odds that person A speaks the truth are 3 : 2 and the odds that person B speaks the truth are 5 : 3. In what percentage of cases are they likely to contradict each other in stating the same fact?

[Ans: 47.5% of cases]

8) In an examination, 30% of the students have failed in Engineering Mechanics, 20% of the students have failed in Mathematics and 10% have failed in both the subjects. A student is selected at random.
   i) What is the probability that the student has failed in Engineering Mechanics if it is known that he has failed in Mathematics?
   ii) What is the probability that the student has failed either in Engineering Mechanics or in Mathematics?

[Ans: (i) $\frac{1}{2}$, (ii) 0.40]

9) Villages A, B, C and D are connected by overhead telephone lines joining AB, AC, BC, BD and CD. As a result of severe gales, there is a probability p (the same for each link) that any particular link is broken. Find the probability of making a telephone call from A to B.

[Ans: $1 - 2p^2 + p^3$]

10) An unbiased coin is tossed four times in succession and a man scores 2 or 1 according to the coin as shows head or tail in each throw. E_1, E_2 are the following events: E_1 – total score is even, E_2 – total score is divisible by 3. Determine whether E_1 and E_2 are independent or not.

[Ans: Not independent]

CHAPTER 7

Sampling and Hypothesis Testing

7.1. SAMPLING

7.1.1. Introduction

Sampling is a process in which the fixed numbers of observations are taken randomly from a larger population.

A technique which is fundamental for behavioural research is known as sampling and without using it, research work is not possible. It is impracticable and impossible to study the whole population due to practical limitations of cost, time and other factors that are indispensable and operative in studying the whole population. For the sole reason of making the research findings economical and accurate, the concept of sampling has been introduced. **For example**, for taking decision about to purchase or not to purchase the fruits, a fruit merchant inspects randomly only a few of them rather than examining each and every fruit.

The important objective of sampling is to obtain maximum information of the population under study using minimum of money, labour and time. A definite plan to obtain a sample from the sampling frame is **sample design**. The method or technique which is adopted by the researcher in selecting the units of sampling from the population is called sampling design.

According to Cocharn, "In every branch of science we lack the resources, to study more than a fragment of the phenomena that might advance our knowledge". In this definition a 'fragment' is the sample and 'phenomena' is the 'population'.

According to Davis S. Fox "in the social science, it is not possible to collect data from every respondent relevant to our study but only from some fractional part is called sampling".

The basis for selecting a sample survey is the framework or roadmap which is called sample design and it affects other important aspects of the survey. For obtaining some type of relevant information using survey, researchers execute it for some population, or universe. The sampling frame must be defined in such a way that it represents the population of interest out of which a sample is to be drawn at random. This sampling frame should be either identical to the population or may be a part of it.

The sampling frame can be indirectly related to the population subject to some under coverage (e.g., the population is preschool children and the frame is a listing of preschools). The design of the sample may be simple or complex which provides the fundamental plan and methodology for the selection of the sample.

7.1.2. Characteristics of a Good Sample Design

The various characteristics of a good sample design are as follows:

1) **True Representative:** The true representative of the population and matching its properties is termed as good sample where aggregate of certain properties is the population and sample is the sub-total of the universe.

Scanned with CamScanner

2) **Free from Bias:** A good sample does not allows prejudices, pre-conceptions, and imaginations which affects its choice and it is unbiased.
3) **Accurate:** A sample is called good when it yields accurate estimates or statistics and free from errors.
4) **Comprehensive:** A sample that is true representative of the population is also comprehensive in nature which is controlled by definite purpose of investigation. A sample's characteristic may be comprehensive but it may not be a good representative of the population.
5) **Approachable:** The subjects of good sample are easily accessible where the tools of research are easily conducted and easy collection of data is possible.
6) **Good Size:** The size of good sample is such that it yields an accurate result and the error due to probability can be estimated.
7) **Feasible:** A good sample creates the research work more feasible.
8) **Goal Orientation:** Any sample which is selected by the researcher should be able to satisfy the objectives of the research. The sample should be taken in proper number. It should be customized to fit the environment under which the research is going to be conducted. If the sample is changed as per the requirement of the survey design then it can come out with better results and outcomes.
9) **Practical:** It means that the concepts of sample selection should be applied properly while conducting the research. The researcher should be well experienced and well instructed. The instructions which are passed to observer should be clear, complete and correct in all terms so as to avoid errors and biasness on their part. The sample should be selected on basis of the sample design. The sampling units should be representative. The sample design should be practical and feasible in nature.
10) **Economical:** It refers that the research should not incur huge costs, time or efforts. One of the objectives of any research is to complete the research with minimum effort, time, money and resources. The researcher calculates per unit cost for each respondent. The researcher should choose that sampling design which gives minimum per respondent cost and maximum accuracy. The funding agency is highly concerned with per unit cost and the corresponding accuracy.

7.1.3. Population

Population or **universe** is the sum total of the objects under study, both animated and in animated. Thus, numerical details of the collection of the individuals or of their qualities, or of results of the operations are known as universe. Universe can be classified as follows:

1) **Finite Universe:** A 'finite universe' comprises of finite number of entities or members. **For example,** the universe of the heights of students in a particular class or the universe of smokers in the Lucknow district.
2) **Infinite Universe:** If the infinite numbers of members are included in a universe then it is called "infinite universe". **For example,** the universe of pressures at various points in the atmosphere.

7.1.3.1. Target Population

The entire group of units, being the focus of study, is represented by the population or universe consisting of all the persons in the country, in a particular geographical location, or a special domestic or economic group based on the purpose and scope of study. Non-human entities like farms, houses or business establishments also comprise population.

For example, if a study is to be conducted on the marks obtained by the students of Statistics then the 'universe' will be all the students of that subject in that class. If the class is consisting of 50 students then the 50 students is the 'universe'.

7.1.3.2. Statistical Population

To draw out statistical inferences from the set of entities of statistical population is based on random sampling of the population. **For example**, to derive generalisations about the crows, to only the set of crows should be the interest of study. However, if one chooses a population such as 'all crows', then the limitations will be in observing crows both the existing as well as existing in the future. There will be additional limitations of geography where the resources are limited for studying crows.

The total of quantities that are measurable or set of numbers is 'statistical population'. Thus, when only a single character symbolised every element of such a set then it is defined as 'single population', **e.g.,** income of individuals. Depending on the finite and infinite number of elements, statistical population can be infinite as well as finite. A random set may not be a statistical population. **For example**, a set of crows at a point of time will not represent a statistical population.

7.1.4. Sample

A part of the population which is examined to estimate the characteristics of the population is called a sample.

For example,

1) A portion of rice is examined to assess the quality of rice in the bag where the portion of rice examined is the sample and the bag of rice is the population.
2) From a large consignment, proportion of defective articles is estimated only from a portion of the consignment. The selected portion is called sample.

7.1.4.1. Sample Frame

The actual set of units and a list containing every member of the population from which a sample is drawn at random is called **sample frame. For example,** if one wants to know the factors that are responsible for the patients being admitted in the hospital after an acute attack of asthma in a specified area, then it becomes necessary to know the names of all those people of that particular area being admitted into the hospital for that reason.

A good sampling frame should be:

1) **Relevant:** Research topic must be directly linked with relevant things.
2) **Complete:** Coverage of all relevant items.
3) **Precise:** It should exclude all the irrelevant items.
4) **Up-to-Date:** Incorporating recent changes and additions, and deleting redundant items.

7.1.4.2. Sample Unit

Every sample is made up of several members or components. These members or components are called **sampling units.** A sampling unit may or may not be a sampling element. Generally sampling unit is referred as sampling element. But in case of sampling units consisting of several population units, each sampling unit is called a cluster of units. The individual population units which come under the cluster of units, are called sampling elements. Therefore in this case a sampling element is an individual object on which the research is done, and it is a component of sampling unit.

For example, in case of inspecting the quality of oranges, only few oranges are selected to examine and each selected orange is a sampling unit. But in case of finding the literacy level of a particular colony, several houses are selected in that colony for research. These houses are sampling units. However the individual members of each house are approached for examining the literacy level, thus these individual members of several

Scanned with CamScanner

7.1.5. Essence of Sampling

An essence of sampling is the process of selection of a part (sample) from the whole (population) for making the inference about the whole. The essence of sampling means that a large number of items (statistical population) may, in definite limits of statistical probability and can presented by a smaller group of items (a sample) chosen from the larger group (a parent population). When a limited selection of items is made from a huge population, then it is known as **sample**. A **limited sample** is usually sufficient to make a generalisation of whole population. In numerous cases, a numbers of persons in population, **for example**, an average yield of entire plots of an agricultural region or pebbles on a sea-beach are so numerous that measuring everything of them would be not possible practically.

In geographical research, an appropriate sampling is extremely necessary as it saves time and cost as well as provides reliable result. This outcome can be used for the purpose of generalization and forecasting.

'Larger the size of sample, the more likely it provide better picture of the parent population' is the simplest rule. It can be said that "the size of sample must be minimum 5% to 15% of the total for satisfactory results. Conclusions on defining parent population and selecting the top sampling technique, though, depend to a big extent on common sense.

7.1.6. Sampling Process

A brief description of which measurements will be taken at what times, on which material, in what manner, and by whom, is termed as 'sampling plan'. The designing of the sampling plans should incorporate that data to be evaluated which is a representative sample of the parameters of interest and be answerable to all the questions as stated in the goals. Following are the steps involved in developing a sampling plan:

1) **Define the Universe:** The universe can be restricted to some geographical limits, a particular type of product or some other constraints. Defining the universe is the initial problem in procedures of sampling. The collection of elements or objects possessing information desired by the researcher regarding what inferences are to be made is termed as the target population or the universe. Precise definition of the target population is must as ambiguous definition will result in ineffective research and deceptive in meaning. Proper defining of the target population requires translating the problem definition precisely as what should and should not be included in the sample.
2) **Sample Frame:** The frame is enlisted by some population that already exists or it is constructed for some study to be conducted by the researcher. After the specification of the population to be studied, the frame of the population is developed. **Sampling frame** consists of all the sampling units of the population which is built by the researcher for the subject of his study or it consists of already existing list of the population.
3) **Specifying the Sampling Units:** A sampling unit is decided on the basis of the sampling frame which is the primary unit containing the sampling elements of the population like, city blocks, households, a business organisation, etc. The design of the project is the partial basis of the selection of the sampling unit. The units that are served as initial sampling are called as **primary sampling units**. The objective of the inquiry is the basis of the composition of one or more units of the population.
4) **Selection of Sample Design:** The procedure of selecting units in the sample involves two primary methods of sampling namely, **probability and non-probability methods** that can be further sub-divided into specific methods of selection. The probability sample being one and the procedure involved in selecting the units has some specific chance of its inclusion in the sample. An arbitrary method is adopted in selecting the units which is not depending upon the chance for the non-probability sample. The purpose of inquiry and the attitude or convenience of the investigators is the basis of this method.
5) **Determination of Sample Size:** The desired degree of accuracy in investigation is directly proportional to the size of the sample which also depends upon the nature of the population and the method of selection. The ideal sample size in marketing research investigations depends upon the type of series and the population size. There will be more degree of heterogeneity and larger units will be drawn in the sample due to larger size of the population. Hence the size of the sample should be larger for being fully representative.
6) **Select the Sample:** Execution of the actual sampling process is called as selection of the sample and the real selection of the elements of sample requiring considerable amount of office and field work especially when the personal interviews are involved. Detailed specification of the sampling design is the requirement for the execution of the sampling process with respect to the implementation of the size of population, sampling frame, sampling unit, sampling technique, and sample size.

7.1.7. Advantages of Sampling

Following are the advantages of sampling:

1) **Saves Time, Money and Effort:** The subjects involved in sampling are small in number which gives him less time to calculate, tabulate, present, analyse, and interpret and hence the researcher saves time, money and effort.
2) **More Effective:** Due to the size of the sample being small than that of the population, the tiredness is reduced in collecting the information and investigator works more effectively.
3) **Faster and Cheaper:** Due to small sample size the data collection, tabulation, presentation, analysis, and interpretation takes less time and it involves nominal expenses.
4) **More Accurate:** Small errors are in sampling because small data are involved in collection, tabulation, presentation, analysis and interpretation.
5) **Gives More Comprehensive Information:** Thorough investigation of the study is the result of small sample which provides more complete information as all the members of the population have equal chance to be included in the sample.

7.1.8. Disadvantages of Sampling

Following are the disadvantages of sampling:

1) **Biased Selection:** Biased selection of the respondents may be included in sampling by the research worker.
2) **Difficulty in Selection:** It is very difficult to select a truly representative sample because a large number of factors obstructing the method of selecting good samples.
3) **Specialised Knowledge Needed:** Specialised knowledge is required in the sampling method where the investigator may commit serious mistakes due to its absence.
4) **Problem of Cooperation:** Due to scattered sample the subjects are uncooperative with the researcher.
5) **Less Accuracy:** When the higher standard of accuracy is expected then the sampling is not suitable.
6) **Limited Nature:** Due to small or dissimilar universe it is impossible to derive a representative sample where census study is the best possible substitute.

Scanned with CamScanner

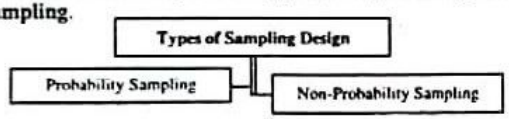
7.2. TYPES OF SAMPLING

7.2.1. Introduction

The plan, method, or technique through which a researcher identifies and selects the potential sampling units from the sampling frame or the target population, to form a relevant sample for the study is termed as **"sample design"**. Sample designing is the working principle of any research process. Without proper designing of sample it is not possible to start a survey, as it is the base which provides the responses of relevant members of the population. The designing of sample may be simple or complex depending on the method used for it.

The two factors namely, the element selection technique and the representation basis are responsible for different types of sample technique. The sample, based on representation, may be either probability sampling (random sampling) or non-probability sampling (non-random sampling). On the basis of element selection the sample can be restrictive or unrestrictive sampling. When each sample element is selected individually from the large population then it is termed as 'unrestrictive sample' and all the other types of sampling are termed as 'restrictive sampling'.

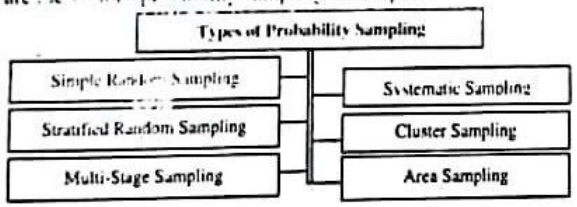
Thus, sample designs are basically of two types, viz., **probability sampling** and **non-probability sampling**.



7.2.2. Probability Sampling

The method in which all units of the universe is given equal chance of being selected in the sample, is known as Probability Sampling. There is an assurance of the results in terms of probability that are obtained through probability or random sampling. The significance of the results lies in measuring the errors of estimation obtained from a random sample which brings predominance of the random sampling designs over the intentional sample design.

Following are the various probability sampling techniques:



7.2.2.1. Simple Random Sampling

This is the most famous and simple method of sampling where each unit of the population is equally probable of getting included in the sample. Let us consider that the size of the population is 'N' from which n units are to be selected at random for a sample such that $^{N}C_{n}$ sample has the probability of being selected equally. Simple random sampling says that:

1) There is an equal chance for each element of the population to be included in the sample and the choices are independent to each other.
2) Each possible sample combination has an equal chance of being chosen.

The possibility of bias or personal prejudices in the selection of the units is eliminated by the method of simple random sampling.

Methods of Simple Random Sampling

Some of the common methods of drawing simple random samples are:

1) **Lottery Method:** This method includes the following steps. Let N be the size of population and a sample of size n is to be drawn:
   i) The units of the population are numbered from 1 to N.
   ii) N identical pieces of papers, cards, capsules or balls are taken and are numbered from 1 to N.
   iii) These are mixed carefully in a bag, bowl or some other container and n items are selected either by replacement or without replacement in one stroke.
   iv) The number of units in the population bearing the numbers on the items drawn in step (iii) constitutes the desired random sample.

   The application of this method is simple and easy but as the size of the population increases to infinity this method is unfeasible.

2) **By Using Random Numbers:** The random number table that helps in selecting a sample has been constructed by some experts. The most popular of all the existing tables, **Tippett's Tables** is very popular and in use. Out of the given population, random selection of the samples is done through these numbers. This table ranges from 0 to 9 having an equal chance of appearing in any position of the table. The actual utilisation of the random numbers is for large population where the random number table consist of several columns and rows out of which one is selected at random and then the samples are continuously selected in sequence for the desired size of the samples. Without any bias it provides a set of random numbers.

7.2.2.2. Systematic Sampling

After the selection of one unit at random from the universe the other units are selected systematically at a specified interval of time.

This method is applicable when the size of the population is finite and on the basis of any system the units of the universe are arranged such as alphabetic arrangement, numerical arrangement or geographical arrangements.

7.2.2.3. Stratified Random Sampling

In the Stratified Random Sampling, the sample is selected from different homogeneous strata or parts of a universe instead of heterogeneous universe as a whole. The summary of this sampling procedure is as follows:

1) The sampled universe is divided (or stratified) into groups that are mutually exclusive and include all items in the universe.
2) A simple random sample is then chosen independently from each group or stratum.

There is a difference in the process of stratified random sampling from that of simple random sampling. In simple random sampling the sample items are chosen at random from the whole universe whereas in stratified sampling the design of the sample is the selection of the separate random sample from each stratum. The distribution of the sample among strata is based on chance in simple random sampling.

Formally, divide the population into non-overlapping groups (i.e., strata)

$N_1, N_2, \ldots\ldots\ldots N_t$

Such that

$N_1 + N_2 + \ldots\ldots\ldots + N_t = N$

Scanned with CamScanner

Then randomly choosing a sample of f =n/N in each stratum, where f is the sampling fraction. **For example,** suppose that the retail sale of a product such as sugar in a universe of 1,10,000 grocery stores is to be studied by the researcher. He will first sub-divide this universe into three strata, based on store size, as stated below:

Table: 7.1

Store Size Stratum	Number of Stores	Percentage of Stores
Large stores	22,000	20
Medium stores	33,000	30
Small stores	55,000	50
Total	**1,10,000**	**100**

Then, by independent random sampling within each of the three strata, the researcher guarantees for the desired sample allocation of stores within each size group, rather leaving their representation to chance. **For example,** according to the stratification scheme it was required to have total samples of 120 stores, the researcher applies, simple random sampling from each strata, that is expected to yield about 24 large stores (20% of 120); about 36 medium stores (30% of 120); and about 60 small stores (50% of 120). Thus all the strata would get representation in the strata. In case of simple random sampling, the equal chances of occurrences are given to all sample points, it may possible that one of the characteristic of small strata may not at all get any representation.

7.2.2.4. Cluster Sampling

According to this method there is further noticeable sub-division of the universe into clusters. Simple random sampling is performed and clusters are drawn accordingly constituting a sample of all the units belonging to the selected clusters. **For example,** if we have to conduct a survey in the city of Mumbai, then the city may be divided into, say, 40 blocks and out of these 40 blocks, 5 blocks can be picked up by random sampling and the people in these five blocks are interviewed to give their opinion on a particular issue. The clusters chosen should be small in size, i.e., more or less the same number of sample units should be there in each cluster.

This method is used in the collection of data about some common traits of the population. Comparing it with random sampling cluster, sampling is certainly less precise and there is no doubt about it in the reduction of the cost by intensifying surveys in selected clusters. Comparatively, within a cluster there is not much information in 'n' observations that appears to be in 'n' randomly drawn observations. Due to its economic advantages cluster sampling is used. Cluster sample estimates are most reliable and effective per unit cost.

Difference between Stratified and Cluster Sampling

Table 7.2 shows the difference between stratified and cluster sampling:

Table 7.2: Difference between Stratified and Cluster Sampling

Stratified Sampling	Cluster Sampling
One divide the population into a few sub-groups: i) There are many elements in each sub-group ii) Selection of the sub-groups depending upon the criterion related to the variables under study.	One divide the population into many sub-groups: i) There are few elements in each sub-group. ii) According to some criterion of ease Sub-groups are selected or by availability in data collection.
Homogeneity is secured within sub-groups.	Heterogeneity is secured within sub-groups.
Securing heterogeneity between sub-groups.	Securing homogeneity between sub-groups.
Randomly choosing the elements within each sub-group.	Randomly choosing the several sub-groups for in-depth typical study.

7.2.2.5. Multi-Stage Sampling

Modification of cluster sampling is multi-stage sampling where in cluster sampling, a sample is constituted by all the units selected in a cluster but in multi-stage sampling the selection of the sample units is in two, three, four stages. Firstly, the universe is divided into first-stage sample units and then further sub-divided into second stage units out of which another sample is selected and similarly third and fourth stage sampling is performed in the same way.

For example, in an urban survey first stage sampling will be of selection of towns and then for each selected towns households will be taken as sub-sample and depending upon the need, individuals will be selected as samples of third-stage sampling.

7.2.2.6. Area Sampling

The area sampling is a form of multi-stage sampling in which maps instead of lists or registers is used as 'sampling frame'. It is commonly used in those countries which do not have proper sampling frame like a population list. For geographic sub-divisions, 'clusters sampling' is the other name of 'area sampling'.

The cluster of units based on geographical area is the primary sampling units known as 'cluster designs' which are famous as area sampling. The positive and negative features of cluster sampling are also applicable to area sampling. The total area for sampling is sub-divided into several smaller areas within which a random sample is selected.

For example, consider a city map, which is used for area sampling. City map has various blocks which are known as 'frame' in case of area sampling and each frame has a unique number for its identification.

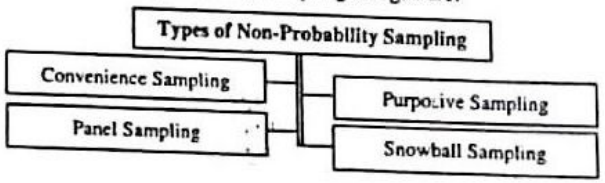
For this sampling (which is based on the geographical considerations) block stratification is employed. Sample of dwellings is taken by identifying the blocks. At the end, each block is subdivided into segments (more or less equal size) and these 'segment samples' may be selected in the sample.

7.2.3. Non-Probability Sampling

Non-probability sampling is that type of sampling procedure which does not have any ground for estimating the probability that whether or not each item in the population has been included in the sample.

There are different names of non-probability sampling such as deliberate sampling, purposive sampling and judgement sampling. In this type of sampling, the researcher deliberately selects items for the sample and the choice of researcher regarding the item is provided more weightage.

In other words, under non-probability sampling the organiser of the inquiry purposively chooses specific units of the universe to constitute a sample on the basis that the small portion selected by him, out of a huge one is typical or representative of the whole universe. The various non- probability sampling designs are:



Scanned with CamScanner

7.2.3.1. Convenience Sampling

On the basis of convenience and approachability, the choice of the sampling units by the researcher, is known as 'convenience sampling'. Samples that are selected accidentally are called 'accidental samples'. Because of the selection procedure (units are selected from their actual place) it is called as 'sample of the man in the street'. Due to their accessibility, samples units are selected. **For example**, by adding the new product in the nearby suitable shops, the potential of the product is tested. This is accomplished by observing the purchasing and selling report of the product.

7.2.3.2. Purposive Sampling

A non-probability sample which follows certain norms is called purposive sampling. Purposive sampling is basically of two types:

1) **Judgement Sampling:** The study which is based on the parameters of population, where the units are selected by a researcher or some other expert on his /her judgement, is called 'judgement sampling'. This technique of sampling is appropriate in the situation where the study of the population is difficult to locate or there are members who are comparatively better than others for an interview in terms of knowledge or interest.

   For example, a sample of grocery stores, considered to be "representative", is selected by a group of sales manager. Experimentally this was found to be a failure because there is no unbiased way of evaluating the precision of sample results. In spite of these limitations, when the size of the sample is extremely small this method is useful.

2) **Quota Sampling:** Quota sampling is the most commonly used non-probability sample designs, which is most comprehensively used in consumer surveys. Principle of stratification is also used by this sampling method. In stratified random sampling the researcher begins by building strata. The common bases for stratification in consumer surveys are demographic, e.g., age, gender, income and so on. Compound stratification is generally used, e.g., gender-wise age groups.

   For each stratum sample sizes are fixed. Regarding stratified random sampling, the sampling within strata may be either proportional or disproportional. Interviews are conducted with the designated quotas by the field workers, with the identification of individual respondents being left to the field-workers.

7.2.3.3. Panel Sampling

In panel sampling a group of participants are selected initially by random sampling method and the same group is asked for the same information repeated number of times during that period of time.

This sample is semi-permanent where members are included repeatedly for iterative studies. In this sampling, there is a facility of selecting and contacting samples that fit in getting high response rate, even by mail.

7.2.3.4. Snowball Sampling

When the characteristic of the desired sample is limited then the special non-probability method is applicable. In this method, it is difficult to locate the respondents because it will be very costly. Depending on the referrals of the initial subjects snowball sampling generates additional subjects.

Though this technique is biased and unable to represent a good cross-section from the population but dramatically it reduces the search cost.

7.3. SAMPLING DISTRIBUTION

7.3.1. Meaning of Sampling Distribution

For a given statistic, the repeated sampling of the population is a sampling distribution which is the probability distribution in statistics. (The calculation of the numerical quantity from the data values in a sample.)

Considering the statistic and the used sample size, the formula of the sampling distribution is based on the distribution of the population. Hence, for all possible samples of given size not only "under repeated sampling trials" there will be a more accurate formula of the distribution of the statistic. **For example,** let us consider a very large normal population having bell-shaped curve and calculates the sample mean ($\overline{X}$) under the assumption that the each sample that has been taken from the repeated samples of the given size, from the population.

Thus, for different samples we have varying sample means and the distribution of these means is the "sampling distribution of the sample mean". As the population is normal, the distribution will be normal. The central limit theorem states that if the given population is "sufficiently well behaved" but not normal, then the sampling distribution of the sample mean will be approximately normal if and only if the size of the sample is sufficiently large.

Hence, if the sample mean is selected as statistic then the mean of the sampling distribution will be equal to the expected value of the statistic.

$$\mu_{\bar{x}} = \mu$$

The standard error of a sample is defined as the standard deviation of the sampling distribution of the statistic. When the mean is selected as statistic, then standard error is:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Where σ denotes the standard deviation of the population distribution of that quantity and n denotes the size (number of items) in the sample.

To find half measurement error one has to expand the sample size (four times) and it is the most important implication of this formula.

To understand cost-benefit trade-offs this factor has its implication in the designing of the statistical samples, where cost is a factor.

When considering the sample median from the same population then the pattern of sampling distribution will be different and usually it will be not normal, but under certain conditions it may be approximately normal.

7.3.2. Sampling Distribution of Mean ($\overline{X}$)

Sampling distribution of $\overline{X}$ is the probability distribution of all the possible values of the sample mean. On the basis of a particular sample size, sampling distribution of the mean holds more importance. Following are the steps involved in the construction of the sampling distribution of the mean:

1) Possible samples of size n are drawn from a finite population of size N.
2) Means are calculated for each sample of equal size.
3) List the means of all possible samples in one column and list the corresponding frequency of happening of each distinct means in the second column.

Scanned with CamScanner

The properties of Sampling Distribution of the Mean with replacement are as follows:

1) **Expected Value of $\bar{X}$:** If the mean is selected as statistic then the value of mean will be equal to the expected value of any statistic.

$$\mu_{\bar{x}} = \mu$$

2) **Standard Deviation of $\bar{X}$:** The standard error of a sample is defined as the standard deviation of the sampling distribution of the statistic. When the mean is selected as statistic then standard error is:

$$\delta_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Where, σ is the standard deviation of the population distribution of that quantity and n is the size of the sample (number of items).

7.3.3. Sampling Distribution of the Difference between Two Independent Sample Means

When all possible simple random samples of size n_1 are drawn from one population and size n_2 samples drawn from a second population, whose parameters are μ_1, σ_1, μ_2 and σ_2, then it is described as the sampling distribution of the difference between two independent sample means. The units of two populations should be independent which means that there should not be any correlation between the samples.

Characteristics of the Difference between Two Independent Sample Means

1) The mean of the difference between sample means $\bar{x}_1 - \bar{x}_2$ is

$$\mu_{\bar{x}_1 - \bar{x}_2} = \mu_1 - \mu_2$$

2) The standard deviation of the difference between sample means $\bar{x}_1 - \bar{x}_2$ is

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}.$$

The value of $\sigma_{\bar{x}_1 - \bar{x}_2}$ with single samples assumes sampling with replacement. When the samples are chosen from the population without replacement then the formula for $\sigma_{\bar{x}_1 - \bar{x}_2}$ is logically accurate in the following conditions:

1) Sample sizes are not more than 10% of the size of the population.
2) Available data is selected from the completely randomised experiment.

The central limit theorem is the basis of the shape of the sampling distribution of the difference in sample means $(\bar{x}_1 - \bar{x}_2)$. The shape of the sampling distribution of $\bar{x}_1 - \bar{x}_2$ is moving towards being normally distributed in the situation when the sizes of the two independent simple random samples are increased. There are less chances of skewness or abnormality in the population distributions because of having two independent samples and the distribution of $\bar{x}_1 - \bar{x}_2$ can be designed strongly by the normal distribution for small samples too.

7.3.4. Sampling Distribution of Sample Proportion ($\bar{p}$)

The classification of each element of the population is possible under various situations into two mutually exclusive categories like, success or failures, accept or reject, head or tail of a coin, and so on. When the sampling procedure has been conducted in an appropriate manner then the similar situations provide the practical examples of binomial experiments.

The sample proportion $\bar{p}$ is the best statistic to be utilised for the statistical inferences about the population parameter p, provided that a random sample of size n is chosen from the binomial population in which x of these has specified characteristics. The sample proportion can be defined as:

$$\bar{p} = \frac{\text{Elements of sample having characteristic (x)}}{\text{Sample size (n)}}$$

Characteristics of Sample Proportion

1) p (i.e., the population proportion) is the expected value or mean of the sampling distribution. Thus, $E(\hat{p}) = p$

2) The standard deviation of p or $\sigma_p = \sqrt{\frac{p(1-p)}{n}}$.

3) We can replace this (σ_p) with z. Hence under the assumption that we have a normal distribution with an excessively large sample size, then our Z becomes:

$$z = \frac{\bar{p} - p}{\sigma_p}$$

So if we were given that p = 0.60 and n = 36 and wanted to know the $P(\bar{p} < 0.53)$.

We can calculate our z = (0.53 – 0.60) / 0.0816 = – 0.86

where $\sigma_p = \sqrt{\frac{0.6(1-0.6)}{36}} = 0.0816$.

7.3.5. Sampling Distribution of the Difference between Two Independent Sample Proportion

When all possible simple random samples of size n_1 are being drawn from one population and samples of size n_2 are drawn from second population then the sampling distribution of the difference between two independent sample proportions $\hat{p}_1 - \hat{p}_2$ is described where the probabilities of a single success in each population are p_1 and p_2. Thus, the samples must be independent and there should be no correlation between units of the two populations.

The differential value in sample proportions $(\hat{p}_1 - \hat{p}_2)$ is $\hat{p}_1 - \hat{p}_2 = \frac{X_1}{n_1} - \frac{X_2}{n_2}$.

Characteristics of the Difference between Two Independent Sample Proportions

Characteristics of the difference between two Independent sample proportions are as follows:

1) The mean of the difference between sample proportions $\hat{p}_1 - \hat{p}_2$ is $\mu_{\hat{p}_1 - \hat{p}_2} = p_1 - p_2$

2) The standard deviation of the difference between sample proportions:

$$\hat{p}_1 - \hat{p}_2 \text{ is } \sigma_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

For a proportion of a single sample it is being assumed that the probabilities of a success in each population will remain constant during the sampling process for the value of $\sigma_{\hat{p}_1 - \hat{p}_2}$. The formula is rationally accurate when the sampling is done without replacement from a population and both the samples are not more than 10% of their respective population sizes. If the conditions $n_1 p_1 \geq 10$, $n_1 (1 - p_1) \geq 10$, $n_2 p_2 \geq 10$, and $n_2 (1 - p_2) \geq 10$ are satisfied then the sampling distribution of the difference in sample proportions can be designed with a normal distribution.

Scanned with CamScanner

7.4. PHARMACEUTICAL NUMERICALS

Example 1: Determine the mean and standard error of the sampling distribution of mean of a random sample of size 25 with replacement from a population of 95 with a mean of 66 and standard deviation of 6.

Solution: By properties 1 and 2 of the sampling distribution of mean, we obtain:

$$\mu_{\bar{x}} = \mu = 66$$

$$\sigma_{\bar{x}} = \frac{8}{\sqrt{25}} = \frac{8}{5} = 1.6$$

Properties of the sampling distribution of the mean without replacement can be summarised as follows:

1) The sampling distribution of mean is the same as the population mean, i.e., $\mu_{\bar{x}} = \mu$.
2) The standard error of the sampling distribution of mean is equal to the standard deviation of the population divided by the square of the sample size multiplied by the square root of the finite population correction factor.

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$$

3) The sampling distribution of mean is approximately normally distributed if the population is large.

Example 2: In a large high school of 2,500 students, the mean number of cars owned by students' families is 2.35 with a standard deviation of 1.06. A second large high school has 2,350 students. The mean number of cars owned by students' families is 2.01 with standard deviation of 1.19.

Simple random samples of 36 students are taken from each school, and the mean number of cars per family is calculated for each. What is the likelihood that the survey will indicate that the average number of cars owned by families in the second school is greater than the first?

Solution: Since both samples are no more than 10% of the population, the formula for the standard deviation of $\bar{x}_1 - \bar{x}_2$ will work. Unless either or both of the populations are extremely skewed, the sample sizes of 36 are sufficiently large to allow the normal distribution to model the sampling distribution of $\bar{x}_1 - \bar{x}_2$.

The mean of the difference in sample means $\bar{x}_1 - \bar{x}_2$ is $\mu_{\bar{x}_1 - \bar{x}_2} = \mu_1 - \mu_2 = 2.35 - 2.01 = 0.34$ cars, and the standard deviation is

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{1.06^2}{36} + \frac{1.19^2}{36}} = 0.266 \text{ cars}$$

The question asks the probability that $\bar{x}_1 < \bar{x}_2$ or $\bar{x}_1 - \bar{x}_2 < 0$. The value of $\bar{x}_1 - \bar{x}_2 = 0$ has a z-score of $z = \frac{0 - 0.34}{0.266} = -1.28$.

The probability of a z-score being −1.28 or less is 0.1003. Thus, the chance that the survey would show a greater average number of cars per family in the second school than the first is about 10%.

Example 3: A tossed tack lands "point up" with probability 0.46. If the tack is tossed 60 times, what are the mean and standard deviation of $\hat{p}$, the proportion of times the tack lands "point up"?

Solution: The probability of landing "point up" is 0.46; this is considered the population proportion p = 0.46. The number of tosses is the sample size n = 60. The mean of the sample proportion $\bar{p}$ is $\mu_{\bar{p}} = p = 0.46$ and the standard deviation is

$$\sigma_{\bar{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.46(0.54)}{60}} = 0.064$$

The distribution of sample proportions is closely related to the binomial distribution. In fact, the mean and standard deviation of the sample proportion $\bar{p}$ are simply the mean and standard deviation of the binomial random variable X when divided by the sample size n. Because of this, the sampling distribution of sample proportions can be modelled with a normal distribution like the binomial distribution if $np \geq 10$ and $n(1-p) \geq 10$.

7.5. HYPOTHESIS

7.5.1. Introduction

It is derived from a Greek word "hypotithenai" which means "to put under" or "to suppose". Hypothesis is the process of taking and using a sample statistics to make an inference for a population parameter.

According to Prof. Morris Hamburg, "A hypothesis in statistics is simply a quantitative Statement about a population".

According to Palmer O Johnson, "A hypothesis is islands in the uncharted seas of thought to be used as bases for consolidation and recuperation as we advance into the unknown".

Hypothesis generally occurs when we perform inductive reasoning. In inductive reasoning, researcher takes a series of observations to perform a theory. So hypothesis is the first step of any research process. It is very important for any research and formulated in various ways. Its basic function is the prediction of the final result of the investigation.

7.5.2. Qualities of a Good Hypothesis

1) **Clear and Precise:** Hypothesis should be clear and precise. If the hypothesis is not clear and precise, the inferences drawn on its basis cannot be taken as reliable.
2) **Ability to Test:** Hypothesis should be capable of being tested.
3) **State Relationship between Variables:** Hypothesis state relationship between variables, if it happens to be a relational hypothesis.
4) **Limited in Scope:** Hypothesis should be limited in scope and must be specific. A researcher must remember that narrower hypotheses are generally more testable and researcher should develop such hypotheses.
5) **Simple:** Hypothesis should be stated as far as possible in most simple terms so that the same is easily understandable by all concerned. But one must remember that simplicity of hypothesis has nothing to do with its significance.
6) **Consistent:** Hypothesis should be consistent with most known facts i.e., it must be consistent with a substantial body of established facts. In other words, it should be one which judges accept as being the most likely.

Scanned with CamScanner

7) **Logical:** If there is two or more hypothesis derived from the same basic theory they should not contradict each other.

8) **Amenable:** Hypothesis should be amenable to testing within a reasonable time. One should not use even an excellent hypothesis, if the same cannot be tested in reasonable time for one cannot spend a life-time collecting data to test it.

9) **Subjectivity:** Researchers subjectivity or his biased judgment should be eliminated from the hypothesis.

7.6. HYPOTHESIS TESTING

7.6.1. Introduction

Hypothesis testing plays an important role in the application of statistics to real-life problems. Hypothesis testing helps the researchers to make the decisions by using sampled data of unknown distribution of a population and its parameters.

For example, consider a manufacturing company which makes the geyser and the average life of the geyser is 1000 hour when manufactured through old process and the average life is 1500 hours when the geyser is manufactured by the new process. So in this case we can make three hypothesis:

1) New process is better than old process.
2) Old process is better than new process.
3) There is no difference between both the processes.

7.6.2. Purpose of Hypothesis Testing

1) Determine the focus & direction for a research effort.
2) Development of a hypothesis forces the researcher to clearly state the purpose of the research activity.
3) Determine what variables will not be considered in a study, as well as those that will be considered.

7.6.3. Types of Hypothesis

Below are some of the important types of hypothesis

1) **Research Hypothesis:** It is directional hypotheses which specify the projected direction of the result. Using earlier research or theory the direction is implied.

   It is a statement which generated by the researchers when they make a guess on the result of an experiment or research. When one needs to test that a data support the research hypothesis then he/she requires to transform the data into statistical hypothesis.

2) **Statistical Hypothesis:** It is a hypothesis that can be verified statistically. It is not necessary that the statement is logical it may be illogical (if statistic verifies it) then this is statistical hypothesis.

   It is a statistical hypothesis that shows statement with one or more parameters (like mean, percentage, etc.) which are population measure under study.

3) **Substantive Hypothesis:** This hypothesis objective is to providing the answer to the raised questions and answer of the felt problem of a research study in a significant way.

   According to the Best and Kalm, "is a formal affirmative statement prediction a single research outcome, a tentative explanation of the relationship between two or more variables".

According to the Stangor, "a substantive or research hypothesis can be defined as a specific and falsifiable prediction regarding the relationship between or among two or more variables, the research hypothesis states the existence of a relationship between the variables of the interest and the specific direction of that relationship".

7.6.4. Hypothesis Decision Table

These tables are the combination of rows and columns and every row corresponds to a single rule. Columns in the table explain the action and condition for the rules. When every condition is valid then the action of a given rule is performed. **Table 7.3** shows a hypothesis decision table.

Table 7.3: Hypothesis Decision Table

Decision	Null Hypothesis	
Accept H_0	Correct Decision	Type II Error
Reject H_0	Type I Error	Correct Decision

7.6.5. Element of Hypothesis Testing

Following are the two basic elements of hypothesis testing.

1) **Test Statistic:** It is determined from the sample. To test the statistic researchers assume that the hypothesis of population parameter is true and then compare the actual value with the hypothetical value of the parameter. If the difference is small, then the hypothesis is accepted, else rejected.

   A statistic on which the decision can be based, whether to accept or reject a hypothesis, is called **Test Statistic (T.S.).** Following are the examples of test statistic:

   T.S. $$Z = \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}}$$ (for Z-test, for two Samples)

   In case of one sample, test statistic (T.S.) is given as below:

   T.S. $$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$ where, $$\frac{\sigma}{\sqrt{n}}$$

2) **Rejection Region:** The rejection region is a subset of the sample space and is associated with the statistics test. If the observed sample falls only within this set, then the researcher can reject the null hypothesis and choose its alternative.

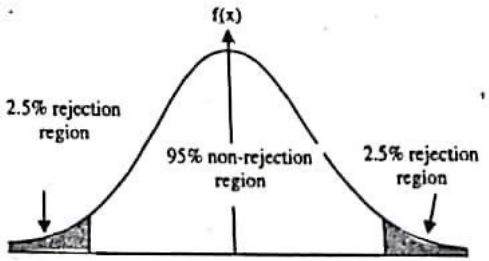


Figure 7.1: Rejection Regions for a Two-Sided 5% Hypothesis Test

Scanned with CamScanner

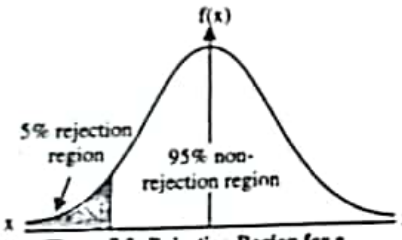


Figure 7.2: Rejection Region for a One-Sided Hypothesis Test of the Form – H_0: $\beta = \beta^*$, H_1: $\beta < \beta^*$

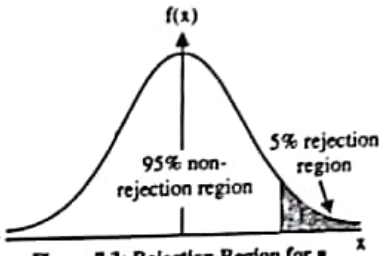


Figure 7.3: Rejection Region for a One-Sided Hypothesis Test of the Form – H_0: $\beta = \beta^*$, H_1: $\beta > \beta^*$

Example 4: 50 smokers were questioned about the number of hours they sleep each day. One wants to test the hypothesis that the smokers need less sleep than the general public which needs an average of 7.7 hours of sleep. The steps below are followed:

i) Compute a rejection region for a significance level of 0.05.

ii) If the sample mean is 7.5 and the population standard deviation is 0.5, what can be concluded?

Solution: The null and alternative hypotheses are as follows:

H_0: $\mu = 7.7$, H_1: $\mu < 7.7$

This is a left-tailed test. The z-score that corresponds to 0.05 is –1.645. The critical region is the area that lies to the left of –1.645. If the z-value is less than –1.645 there one will reject the null hypothesis and accept the alternative hypothesis. If it is greater than –1.645, then one will fail to reject the null hypothesis and say that the test was not statistically significant.



$$Z = \frac{7.5 - 7.7}{0.5/\sqrt{50}} = -2.83$$

Since –2.83 are to the left of –1.645, it is in the critical region. Hence the null hypothesis is rejected and the alternative hypothesis is accepted. One can conclude that smokers need less sleep.

7.6.6. Steps / Procedure of Hypothesis Testing

Step 1) Set up a Hypothesis: To setup the hypothesis, first set the null hypothesis (denoted by H_0). Null hypothesis is tested for the possible rejection under the assumption that it is true. It is expressed in the form of equation.

Null hypothesis is given as follows: H_0: $\mu = \mu_0$

Where, μ = Population Means and

μ_0 = Hypothesized Value of the Population Mean.

For example, to test whether a population mean is equal to 140, the null hypothesis can be set as "population mean is equal to 140". H_0: $\mu = 140$

Now set the alternative hypothesis which is rationally opposite to the null hypothesis and

Alternative hypothesis is given as: H_1: $\mu \neq \mu_0$

Therefore, H_1: $\mu < \mu_0$

H_0: $\mu > \mu_0$

For the above example, the alternative hypothesis can be set as "population mean is not equal to 140".

So, H_1: $\mu \neq 140$

This gives the two other alternative hypotheses:

H_1: $\mu < 140$, Indicates i.e., the population mean is less than 140 and

H_1: $\mu > 140$; Indicates i.e., the population mean is greater than 140.

Step 2) Set up a Suitable Significance Level: Significance level is related to the null hypothesis and denoted by α. The size of the rejection or critical region is also known as the level of significance.

Ensure that the obtained result is free from decision-maker's biasness of choice. For this, the significance level must always be determined before we draw samples.

0.01, 0.05, and 0.10 are the most commonly used levels of significance by researchers.

Step 3) Test Statistic: In this step one can select the suitable statistical test for the analysis. On the basis of the type, number and level of data, statistical test is decided. To select the statistical test, statistics (mean, proportion, etc.,) used in the study is also considered to get the good result.

Step 4) Doing Computations: After selecting the statistical test, the researcher tests the performance of different calculations from a random sample of size n. Here researcher also computes the standard error.

Step 5) Making Decision: In this step, researcher takes the decision that the null hypothesis is to be rejected or accepted. Following conditions are checked to accept and reject a null hypothesis:

i) When the calculated value is less than the critical value then test statistics falls in the acceptance region and null hypothesis is accepted.

ii) When the calculated value is greater than the critical value then test statistics falls in the rejection region and null hypothesis is rejected.

Generally, in the testing of hypothesis 5% level of significance ($\alpha = 0.05$) is used to take the decision.

7.6.7. Advantages of the Tests of Hypothesis

1) It is suitable for establishing the focus and direction in a research effort.
2) It helps the researcher in the development of a hypothesis force which shapes the purpose of the research movement.
3) It determines which variables will not be measured in a study and which will be measured.

7.6.8. Disadvantages of the Tests of Hypothesis

1) The tests should not be used in a mechanical fashion. They should keep in mind that testing is not decision-making itself. The tests are only useful aids for decision-making. Hence "proper interpretation of statistical evidence is important for intelligent decisions."
2) Tests do not explain the reasons as to why the differences exist, say between the means of the two samples. They simply indicate whether the difference is due to fluctuations of sampling or because of other reasons but the tests do not tell us as to

Scanned with CamScanner

3) Results of significant tests are based on probabilities and cannot be expressed with full certainty. When a test shows that a difference is statistically significant, then it simply suggests that the difference is probably not by chance.
4) Statistical inferences based on the significant tests cannot be said to be entirely correct evidences concerning the truth of the hypothesis. This is especially in case of small samples where the probability of drawing erring inferences happens to be generally high. For greater reliability, the size of samples is sufficiently enlarged.

7.7. IMPORTANT TERMS

7.7.1. Null Hypothesis

A statistical hypothesis which is stated for the purpose of possible acceptance is called null hypothesis. It is usually denoted by the symbol H_0. **For example,** the null hypothesis may be expressed symbolically as:

$H_0 : \mu = 162$ cms.

According to Prof R.A. Fisher, "Null hypothesis is the hypothesis-which is tested for possible rejection under the assumption that it is true."

The following may be borne in mind in setting the null hypothesis:

1) If we want to test the significance of the difference between a statistic and the parameter or between two sample statistics then we set up a null hypothesis that's difference is not significant. This means that the difference is just due to fluctuations of sampling:
   $H_0 : \mu = \overline{X}$
2) If we want to test any statement about the population we set up the null hypothesis that it is true. **For example,** it we want to find it the population mean has specified value μ_0, then we setup the null hypothesis $H_0 : \mu = \mu_0$

7.7.2. Alternative Hypothesis

Any hypothesis which is complementary to the null hypothesis is called an alternative hypothesis and is usually denoted by H_1 or H_a. **For example,** if we want to test the null hypothesis that the average height of the soldiers is 162 cms., i.e.,

$H_1 : \mu = 162$ cms. $= \mu_0$ (say)

Then the alternative hypothesis could be:

1) $H_1 : \mu \neq \mu_0$ (i.e., $\mu > \mu_0$ or $\mu < \mu_0$)
2) $H_1 : \mu > \mu_0$
   $H_1 : \mu < \mu_0$

7.7.3. Errors in Hypothesis Testing

Two types of errors occur in hypothesis testing:

1) Type I Error
2) Type II Error

Following possibilities occur during the hypothesis testing.

1) Test leads to rejection while hypothesis is true. (Type I Error)
2) Test leads to acceptance while hypothesis is false. (Type II Error)
3) Test leads to acceptance and hypothesis is also true. (No Error)
4) Test leads to rejection and hypothesis is false. (No Error)

The first two possibilities lead to errors. If we reject a hypothesis when it should be accepted (possibility 1), then the error is known as type I error.

And if we accept a hypothesis when it should be rejected (possibility 2), then the error is known as type II error.

Consider the following table which provide the basic idea of Type I and Type II error.

	Accept H_0	Reject H_0
H_0 is true	No Error	Type I Error
H_0 is false	Type II Error	No Error

7.7.3.1. Type I Error

If we reject correct hypothesis then it comes under Type I error and denoted by α.

So, α = Probability of Type I Error = Probability of rejecting H_0 when H_0 is true.

When the difference of the two means is zero and test indicates the rejection of the null hypothesis, then it is known as Type I Error.

7.7.3.2. Type II Error

If we accept the wrong hypothesis then it comes under the Type II error and denoted by β.

So, β = Probability of Type II Error = Probability of accepting H_0 when H_0 is not true (false). When the difference of the two means is not zero and the test indicates the acceptance of the null hypothesis then it is known as Type II Error.

7.7.4. Level of Significance

We check the validity of the null hypothesis (H_0) against the alternative hypothesis (H_1) at a certain level of significance to setup the hypothesis. Researchers generally use the 5% and 1% level of significance in the research.

Factors Affecting the Level of Significance

1) Size of the difference between sample means.
2) Total size of the samples.
3) Variability of measurements within the samples and
4) Direction (either directional or non-directional) of the hypothesis.

Note: A directional hypothesis predicts the direction of the difference between means and difference between proportions etc.

7.7.5. Degree of Freedom

In statistics, the number of elements in the calculation which are able to vary is known as degree of freedom.

For example, consider that if the mean of the data is 30 and that the values are 20, 15, 45, and one value is unknown. To find the mean of a list of data, add all the data and divide it by the total number of values.

Now, if the unknown value is x then using the mean formula:

$$= \frac{20 + 15 + 45 + x}{4} = 30$$

So, $x = 40$

Scanned with CamScanner

Now take the situation where two values are missing and are denoted by x and y. So using the mean formula we find that,

$$\frac{20+15+45+x+y}{5}=30;\quad \frac{80+x+y}{5}=30;\ x+y=70\ x=(70-y)$$

It shows that, when we select a value for x, then the value for y is determined. This shows that here the degree of freedom is one. For n samples the degree of freedom will be (n – 1).

For example, for the sample size of 15, the degree of freedom will be 14. For the s × t size contingency table, the degree of freedom will be ((s – 1) (t – 1)).

Where s = Number of columns and
t = Number of rows

Now, if the contingency table is of 2×2, then the degree of freedom is (2 – 1) (2 – 1) = 1.

7.7.6. One Tailed and Two Tailed Tests

When the population parameter is known in a hypothesis test and it falls to the right or the left of the centre of the sampling distribution, then it is known as one tailed test.

This test is used to check the increase and decrease in the parameter. So, it is used to test the null hypothesis.

For example, junior class students will not score significantly greater than higher class students in an IQ test; women are not significantly stronger than men. In every case, the null hypothesis predicts the direction of the expected difference. When hypothesis predicts that one value is dissimilar from another value and does not predict which is higher, then such hypothesis is known as two-tailed (non-directional).

7.7.7. Power of Statistical Test/Measuring the Power of Hypothesis Test

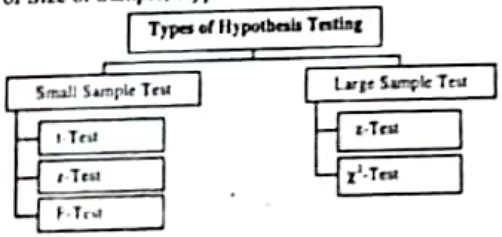
The power of a test denotes the capability of the test to make a right decision, i.e., to reject the null hypothesis in case it is false. The power of a test can also be described as the probability of not making a type 2 error. Mathematically it can be expressed as follows:

$$\text{Power} = 1 - p\,(\text{Type II error}) = 1 - \beta$$

The maximum power a test can have is 1, the minimum is 0. Ideally we want a test to have high power, close to 1. A power curve is the result of plotting 1-β for each possible value of the population parameter. Absolute care should be taken not to reject the null hypothesis (α) when it is correct.

7.8. TYPES OF HYPOTHESIS TESTING

On the Basis of Size of Sample, Hypothesis testing can be categorised as follows:



7.8.1. Small Sample Test

When the sample size (n) is less than 30 (n<30) then it is known as small sample.

According to rule, theory and methods of small samples are appropriate to large samples; but for small samples, the theory and method of large samples are not applicable. In testing a given hypothesis, the small samples are used. It determines the experiential values, which could have risen through sampling variations from certain values given in advance.

For example, in case, a sample of 12 gives a correlation coefficient of 0.5, one can find out whether the value is significant of correlation in the parent population. The investigator's estimation will vary usually from sample to sample in a small sample. A result comes from a small sample is generally less precise than the result drawn from a large sample result.

Types of Small Sample Test

In case of small sample following tests are used:

1) t- Test 2) z-test 3) F-Test

7.8.2. Large Sample Test

When the sample size (n) is greater than 30 (n > 30), then it is known as large sample. Generally large sample is concerned with the variables such as height, weight, etc.

Assumptions for Large Sample Tests

An assumption which are made for a large samples are as given below:

1) In statistics, the random sampling is approximately normal.
2) The values of sampling are sufficiently closer to the value of population and for the calculation of standard error of estimate, it can be used.

Types of Large Sample Test

Following are the most prominent large sample tests:

1) z-test 2) χ^2-Test

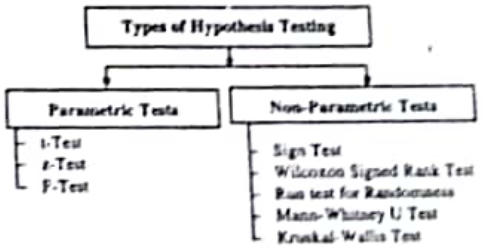
Difference between Large and Small Sample Tests

Table 7.4 shows the difference between large and small sample tests:

Table 7.4: Difference between Large and Small Sample Tests

Basis of Difference	Large Sample	Small Sample
Sample Size	It is greater than 30, i.e., n > 30	It is less 30, i.e., n < 30
Assumption for Test of Significance	The random sampling distribution of a statistic is approximately normal	Assumption of normality is made unless otherwise stated
Comparison from Population	Different values of samples are close to population values.	Sample values and population values differ significantly
Standard Error	This concept is used while testing for significance	Generally, this concept is not used
Tests of Hypothesis	z-test, chi-squared or χ^2 test.	t-test, z-test, F-test.

On the Basis of Parameter Used, the hypothesis testing can be divided into following two categories:



Scanned with CamScanner

7.8.3. Parametric Tests

Parametric tests are based on models with some assumptions. If the assumptions hold well, these tests may offer a more powerful tool for analysis. Parametric tests make certain assumptions about the population from which the samples are taken.

Generally, non-parametric tests make far lesser assumptions and can be utilised for testing different types of data. However, parametric tests are preferred more because non-parametric tests are more susceptible to errors on account of not being less sensitive in identifying variations in samples or in ascertaining the affect that an independent variable may have on a dependent variable.

7.8.4. Non-Parametric Tests

The practical data may be abnormal and/or it may not be possible to estimate the parameter(s) of the data. The tests which are used in such situations are called non-parametric tests. As the data is free from being classified into different parameters, these tests are called non-parametric tests. The level of calculation required in non-parametric tests is quite less as the computation of the parameters is not required.

These tests can be applied on a very small sample size especially while conducting a pilot study. As the interval data is typically converted into ordinal data, there is no need to make any assumption about the parent population. The biggest advantage of using non-parametric tests is that they can handle ordinal level of data.

7.9. STANDARD ERROR OF MEAN(SEM)

7.9.1. Introduction

The standard error of the mean is denoted by symbol σ_M. It is the standard deviation of the sampling distribution of the mean.

Mathematically, it can be represented as:

$$\sigma_M = \frac{\sigma}{\sqrt{N}}$$

Where,
σ = Standard deviation of original distribution
N = Sample size

This formula does not consider a normal distribution. This formula indicates that the larger the size of sample, the smaller the standard error of the mean. The standard error size of the mean is inversely proportional to square root of size of the sample.

7.9.2. Assumptions and Usage

Let us assume that the data is normally distributed. Then for calculating the approximate confidence intervals for the mean, quartiles of the normal distribution and the sample mean and standard error can be used. To determine the upper and lower 95% confidence limits, the following expression can be used:

Upper 95% Limit = $\overline{X} + (S_E 1.96)$

Lower 95% Limit = $\overline{X} - (S_E 1.96)$

Where, $\overline{X}$ = Sample mean
S_E = Standard error for the sample mean

And 1.96 is the 0.975 quantile of normal distribution.

In specific, standard error of a sample statistic is projected standard deviation of error in the process by which it was generated. This means that the sampling distribution of sample statistic is the standard deviation of it.

The simple measures of uncertainty in a value are generally provided by the standard errors and are often used as:

1) In case, standard error of several specific quantities is given then the standard error of some function of the quantities can be calculated easily in several cases;

2) Where the probability distribution of values is given and then it can be helpful to calculate a good approximation to an exact confidence interval.

Example 5: The sample of 900 members has a mean value 3.4 cms and S.D. value 2.61 cms. Find the 95% confidence limits of population mean.

Solution: S.E. of mean $= \frac{2.61}{\sqrt{900}} = \frac{2.61}{30} = 0.087$

For 95% confidence limits, the value of Z is 1.96.

$\therefore$ 95% confidence limits of mean are $3.4 \pm 1.96 \times 0.087$

$\Rightarrow$ The limits are $3.4 - 1.96 \times 0.087$ and $3.4 + 1.96 \times 0.087$ i.e., 3.23 to 3.57.

Example 6: The aircraft manufacture wants to purchase aluminium sheets of thickness 0.05 inches. Thicker sheets would be too heavy and thinner sheets would not be suitable. An aircraft manufacturer takes a random sample of 100 sheets with average width equal to 0.048 and Standard deviation equal to 0.01 inches. Would the aircraft manufacturer purchase the aluminium sheets from his supplier if she needs to make the decision at:
i) 5% level of significance
ii) 1% level of significance?

Solution: Suppose H_0 is the null hypothesis that he would purchase.

H_1 is alternative hypothesis that would not purchase.

Since the manufacturer does not want more deviation on either side, we would apply two tailed test:

$$Z = \frac{\overline{x}_1 - \overline{x}}{\text{S.E of mean}} = \frac{0.048 - 0.05}{\sigma/\sqrt{n}} = \frac{0.002 \times 10}{0.01} = -2$$

A tabulated value of Z at 1% level of significance is 2.58 and at 5% level of significance for two tailed test is 1.96.

Therefore, the manufacturer can reject at 5% level of significance and accept the sheets at 1% level of significance.

Scanned with CamScanner

7.10. EXERCISE

7.10.1. Very Short Answer Type Questions

1) What is sampling?
2) Define the sample with example.
3) What is probability sampling?
4) List the disadvantages of non-probability sampling.
5) Write the short notes panel sampling.
6) Define the snowball sampling.
7) What is sampling distribution?
8) What is hypothesis?
9) Give the purpose of hypothesis testing.
10) State and define standard error of mean.

7.10.2. Short Answer Type Questions

1) What are population, target and statistical population?
2) Write the sample frame and sample unit with example.
3) Give the comparison between the cluster, multi-stage, area sampling.
4) Differentiate between the probability and non- probability sampling.
5) Discuss the qualities of a good hypothesis.
6) Give a definition of null hypothesis and alternative hypothesis.

7.10.3. Long Answer Type Questions

1) Discuss the characteristics of sample and sampling process. Also state the advantages of sampling.
2) Explain the simple random and systematic, stratified random sampling.
3) What is non-probability sampling? And explain the convenience and purposive sampling.
4) What is sampling distribution? Explain its properties and mean.
5) What do you mean by hypothesis testing? And explain the various type of hypothesis.
6) Give a brief explain errors in hypothesis testing and level of significance.
7) Explain the following terms:
   i) Degree of Freedom
   ii) One Tailed and Two Tailed Tests
   iii) Power of Statistical Test

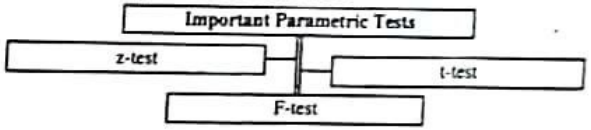
CHAPTER 8 Parametric Tests

8.1. PARAMETRIC TESTS

8.1.1. Introduction

There are some models with assumptions on which parametric tests are based. These tests can be more sophisticated methods for the analysis if the assumptions are fulfilled appropriately. It has some assumptions which are closely related with the traits of the parent population from which the samples are drawn. Although, there is lesser number of assumptions in case of non-parametric tests and these can be employed effectively over wide range of data, yet they are not preferred over parametric tests as they have lesser sensitivity in identifying the differences among various samples or the effect on dependent variable by the independent variable.

Figure below show Important Parametric Tests



8.1.2. Assumptions about Parametric Test

The various assumptions which are associated with the parametric test are as follows:

1) Observations are taken from general population,
2) The size of the sample is large and
3) Population parameters are in good condition.

8.1.3. t–Test (Student's 't' Distribution)

A particular frequency distribution is obtained when a very large number of small samples are taken from a population and the mean of these samples is used to plot the frequency distribution. It is called **Student's t-distribution.**

Sir William Gossett and R.A. Fisher are considered to be the great contributors towards the theory of small samples. Under the name 'students', Gossett published his findings in the year 1905 and thus this theory is commonly known as 'students' distribution', 't-test' or 'students' t-distribution'. The t-distribution can be implemented in situations when the sample size is 30 or less and no information is available with respect to the population standard deviation.

The formula is, $t = \frac{(\bar{X} - \mu)}{s}\sqrt{n}$

Where, Sample Standarad Deviaion(s) $= \sqrt{\frac{\Sigma(X - \bar{X})^2}{n-1}}$

If sample S. D. is given without using n-1 as denominator then

$$t = \frac{(X - \mu)}{s/\sqrt{n-1}}$$

Scanned with CamScanner

8.1.3.1. Properties of t - Distribution

1) Similar to normal distribution, a bell shaped frequency curve is also associated with t-distribution which is symmetrical in nature. Degree of freedom is the only parameter which determines the shape of the curve. The shape of the curve changes with the change in degree of freedom.
2) Similar to normal curve, the mean of t-distribution is zero.
3) The value of variance of t-distribution is usually greater than 1 and as the sample size increases it shifts towards unity. Therefore, t-distribution becomes normal distribution in case of sample size being very large.
4) It can also be used for large samples.
5) The value of t-distribution ranges from -1 to +1.

8.1.3.2. Assumptions for Student's t-Test

1) The parent population from which the sample is drawn is normal.
2) The sample observations are random, i.e., the given sample is drawn by random sampling method.
3) The population standard deviation σ is not known.

8.1.3.3. Applications of T-Distribution

In statistics, t-distribution has a lot of applications, some of them are as follows:

1) For testing, in case, the sample means ($\bar{x}$) differs significantly from the hypothetical value (μ) of the population mean.
2) For testing the significance of difference between two samples means.
3) For testing the significance of a sample regression co-efficient and observed sample correlation co-efficient.
4) For testing the significance of multiple correlation coefficients and the observed partial correlation co-efficient.

8.1.3.4. Graph of 't' Distribution

The figure 5.6 shows the shape of t-distribution and a normal distribution. It is clear that when the degree of freedom changes then the shape of the t-distribution also changes. The above diagram represents two important characteristics of t-distribution which are as follows:

1) A t-distribution is lower at the mean and higher at the tail than a normal distribution.
2) The t-distribution has proportionately greater area in its tail than the normal distribution. Interval widths from t-distributions are therefore, wider than those based on the normal distribution.

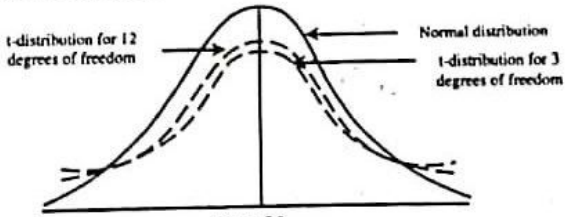


Figure 8.1

8.1.4. Types of T-Test

There are two types of T-tests:

1) Sample t-Test
2) Pooled T-Test/unpaired t-Test
3) Paired t-Test

8.1.4.1. Sample t-Test

If the variance of a normal population is unknown and one wants to determine the mean of a sample which is drawn from the population, and it deviates significantly from a stated valued (the hypothetical value of the population mean), then the following statistic is used:

$$t = \frac{(\bar{X} - \mu)\sqrt{n}}{s}$$

where, $\bar{X}$ = the mean of the sample,

μ = the actual or hypothetical mean of the population

n = the sample size,

s = the standard deviation of the sample.

$$s = \sqrt{\frac{\Sigma(X-\bar{X})^2}{n-1}} \text{ or } s = \sqrt{\frac{\Sigma d^2 - n(d)^2}{n-1}} = \sqrt{\frac{1}{n-1}\left[\Sigma d^2 - \frac{(\Sigma d)^2}{n}\right]}$$

Where, d = deviation from the assumed mean.

Confidence or Fiducial Limits for μ

For given degrees of freedom v and for a given level of confidence α the fiducial limits for μ are given by

$$\bar{X} \pm t_{\alpha}(v) \times \frac{s}{\sqrt{n}},$$

Thus 95, percent confidence limits (or 5 percent level of significance or for $\alpha = .05$) the confidence limits of μ are

$$\bar{X} \pm t_{.05} \times \frac{s}{\sqrt{n}} \text{ and}$$

99 percent confidence limits (or $\alpha = .01$) for μ are

$$\bar{X} \pm t_{.01} \times \frac{s}{\sqrt{n}}$$

Procedure

Step 1: Set up the null hypothesis.

Step 2: Then we find standard deviation from the following formula:

i) When deviation is taken from actual mean

$$s = \sqrt{\frac{\Sigma(X-\bar{X})^2}{n-1}}$$

ii) When deviation is taken from assumed mean

$$s = \sqrt{\frac{\Sigma d^2 - n(d)^2}{(n-1)}} = \sqrt{\frac{1}{(n-1)}\left[\Sigma d^2 - \frac{(\Sigma d)^2}{n}\right]}$$

Step 3: Degree of freedom = v = n – 1

Step 4: Then put this value of standard deviation and find the value of t.

Step 5: If t_{cal} < table value at the significance level then null hypothesis is accepted.
If t_{cal} > table value at the significance level, then null hypothesis is rejected and one accepts alternative hypothesis.

Example 1: The mean weekly sale of soap bars in departmental stores was 146.3 bars per store. After an advertising campaign the mean weekly sales in 22 stores for a typical week increased to 153.7 and showed a standard deviation of 17.2. Was the advertising campaign successful? (Given tabulated value of t for 21 d.f. at 5% level of significance = 1.72)

Scanned with CamScanner

Solution: Given, $n = 22$, $\bar{X} = 153.7$, $s = 17.2$

Null Hypothesis

$H_0 = \mu = 146.3$ bars, i.e., the difference between $\bar{X}$ and μ is not significant. In other words, the advertising campaign is not successful.

Alternative Hypothesis

$H_1: > 146.3$ bars (right-tailed)

Test Statistics

Under the null hypothesis, the test statistic is: $t = \dfrac{\bar{X} - \mu}{s/\sqrt{n-1}}$

Now $t = \dfrac{153.7 - 146.3}{17.2/\sqrt{21}} = \dfrac{7.4\sqrt{21}}{17.2} = 1.9716$

Tabulated value of t for 21 degree of freedom at 5% level of significance level for single-tailed test is 1.72. Since the calculated value is greater than the tabulated value, it is significant and we reject the null hypothesis. Hence the advertising campaign was successful.

Example 2: The random sample of 10 boys had the following I.Q.'s:
70, 120, 110, 101, 88, 83, 95, 98, 107, and 100. Can these data support, the assumption of a population mean I.Q. of 100? Find out, a reasonable range in which most of the mean I.Q. values of samples of 10 boys exist.

Solution: Let us consider that the null hypothesis H_0 shows that the data are consistent with the assumption of a mean I.Q. of 100 in the population, that is $\mu = 100$
And H_1 is the alternative hypothesis, $\mu \neq 100$

Under H_0, the test statistic is: $t = \dfrac{(\bar{x} - \mu)}{\sqrt{\dfrac{S^2}{n}}} \sim t_{(n-1)}$,

Where $\bar{x}$ and S^2 are to be estimated from the sample value of I.Q.'s.

Table 8.1: Calculation for Sample Mean and S.D.

X	$(X - \bar{x})$	$(X - \bar{x})^2$
70	–27.2	739.84
120	22.8	519.84
110	12.8	163.84
101	3.8	14.44
88	–9.2	84.44
83	–14.2	201.64
95	–2.2	4.84
98	0.8	0.64
107	9.8	96.04
100	2.8	7.84
Total 972		1833.60

Therefore, $n = 10$, $\bar{x} = \dfrac{972}{10} = 97.2$ and $S^2 = \dfrac{1833.60}{9} = 203.73$

$\therefore |t| = \dfrac{|97.2 - 100|}{\sqrt{\dfrac{203.73}{10}}} = \dfrac{2.8}{\sqrt{20.37}} = \dfrac{2.8}{4.514} = 0.62$

Tabulated value $t_{0.05}$ for $(10 - 1)$, i.e., 9 degree of freedom (d.f.) for two-tailed test is 2.262.

Conclusion: Because calculated $t < t_{0.05}$ for 9 d.f., H_0 may be accepted at 5% level of significance and we may determine that the data are consistent with the assumption of mean I.Q. of 100 in the population. The 95% of confidence limits in which the mean I.Q. values of samples of 10 boys will exist are given by following formula:

$\bar{x} \pm t_{0.05}\dfrac{S}{\sqrt{n}} = 97.2 \pm 2.262 \times 4.514 = 97.2 \pm 10.21 = 107.41$ and 86.99

Therefore, 95% of confidence interval is [86.99, 107.41].

8.1.4.2. Pooled t-Test/Unpaired t-test

By using a technique known as **Pooled t-Test,** we will compare the means of the two independent samples. The unpaired t-test is statistical technique that compares means of 2 independent or unrelated groups to find whether there is a significant variance between the two. It is also known as **independent t-test.** If the two independent samples of size n_1 and n_2 with means $\bar{X}_1$ and $\bar{X}_2$ and standard deviation S_1 and S_2 are given and one wants to test the hypothesis that the samples come from the same normal population. To carry out the test, we calculate the statistic as follows:

i) The test statistics which is followed in this case is

$$t = \frac{\bar{X}_1 - \bar{X}_2}{S} \times \sqrt{\frac{n_1 n_2}{n_1 + n_2}}$$

Where, $\bar{X}_1$ = mean of the first sample, $\bar{X}_2$ = mean of the second sample
n_1, n_2 = number of observations in the first and second sample;
S = combined standard deviation.

ii) The value of S is calculated by the following formula:

$$S = \sqrt{\frac{\Sigma(X_1 - \bar{X}_1)^2 + \Sigma(X_2 - \bar{X}_2)^2}{n_1 + n_2 - 1}}$$

The deviations should be taken from assumed means when the actual means are in fraction. In such a case the combined standard deviation is obtained by applying the following formula:

$$S = \sqrt{\frac{\Sigma(X_1 - A_1)^2 + \Sigma(X_2 - A_2)^2 - n_1(\bar{X} - A_1)^2 - n_2(\bar{X}_2 - A_2)^2}{n_1 + n_2 - 2}}$$

A_1 = Assumed mean of the first sample, A_2 = Assumed mean of the second sample
$\bar{X}_1$ = Actual mean of the first sample, $\bar{X}_2$ = Actual mean of the second sample
The degrees of freedom = $(n_1 + n_2 - 2)$

iii) When the number of observations and the standard deviation of the two samples are given then the pooled estimate of standard deviation can be obtained as follows:

$$S = \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}$$

iv) If the calculated value of t is greater than the tabulated value at any specified significance level (5% or 1%), the difference between the sample means is said to be significant at specified level of significance otherwise the data is said to be consistent with the hypothesis.

Scanned with CamScanner

Hypotheses of an Unpaired t-Test

1) H_0 (Null Hypothesis): "There is no significant difference between the means of 2 groups."
2) H_1 (Alternative Hypothesis): "There is a significant difference between 2 population means, and this difference is unlikely to be caused by sampling error."

Assumptions for Testing the Difference of Means

1) The universe from which the samples are drawn are normally distributed.
2) The two samples are random samples and have been drawn independently.
3) The variances of the two populations are equal. It follows that the standard deviations of the two universe or σ_1 and σ_2 are also equal.
4) Population variances are not known.

Procedure for Testing the Difference of Means

Step 1: Set up the null hypothesis.

Step 2: Then we find standard deviation from the following formula:

i) When deviation is taken from actual mean

$$S=\sqrt{\frac{\Sigma(X_1-\bar{X}_1)^2+\Sigma(X_2-\bar{X}_2)^2}{n_1+n_2-1}}$$

ii) When deviation is taken from assumed mean

$$S=\sqrt{\frac{\Sigma(X_1-A_1)^2+\Sigma(X_2-A_2)^2-n_1(\bar{X}-A_1)^2-n_2(\bar{X}_2-A_2)^2}{n_1+n_2-2}}$$

iii) When standard deviation of two samples is given

$$S=\sqrt{\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}}$$

Step 3: Degree of freedom $= n_1 + n_2 - 2$

Step 4: Then put this value of standard deviation and find the value of t.

Step 5: If t_{cal} < table value of significance level, then null hypothesis is accepted.
If t_{cal} > table value of significance level, then null hypothesis is rejected and alternative hypothesis is accepted.

Example 3: Suppose that we compared TV watching hours for female and male students of a college and cannot discard the hypothesis that they had the same population means. A sample size for males was only about 1/2 than females, but the standard deviation was larger for males. Following summary is given below:

	Sample Size	Mean	Standard Deviation	Variance
Male	n_1 59	2.37	1.87	3.50
Female	n_2 116	1.95	1.51	2.28

A difference in means is given as 0.42 hours, the unpooled standard error is 0.281, 1.49 is test statistics, and 0.14 is resulting p-value. Examine the data and apply Pooled-Test.

Solution: Here are the computations for the pooled t-test:

$$s_p^2=\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}=2.69,\ s_p=\sqrt{2.69}=1.64$$

$$\text{Pooled s.e. } \bar{x}_1-\bar{x}_2=\sqrt{\frac{s_p^2}{n_1}+\frac{s_p^2}{n_2}}=s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}=1.64\sqrt{\frac{1}{59}+\frac{1}{116}}=0.262$$

$$t=\frac{\text{Sample statistic}-\text{Null value}}{\text{Pooled standard error}}=\frac{0.42}{0.262}=1.60$$

Degree of Freedom(d.f) = 59 + 116 – 2 = 173; p value = 0.11

While the p-value for pooled test still would not lead to rejection of null hypothesis, it is remarkably smaller than the p-value of 0.14 for unpooled test.

Example 4: In a mutation breeding experiment, gamma-irradiation effect was evaluated on 200 seed weight in grams per plant of a mungbean variety in M_2 generation. The experimenter obtained the following results. Analyse the data using the t-test and give your inference as regards the effect of gamma-irradiation.

200-Seed Weight (Grams) per Plant

Control	2.9	3.2	3.5	3.4	3.1	4.0	3.8	3.0	4.1	4.0
Treated	2.9	2.8	3.0	3.5	3.7	3.3	3.0	3.1	2.9	2.8

Solution: Null Hypothesis: Ho: There is no significant difference between the mean 10 seed weight per plant in the control and treatment.

200-Seed Weight (Grams)

Control		Treated	
x_1	$(x_1)^2$	x_2	$(x_2)^2$
2.9	8.41	2.9	8.41
3.2	10.24	2.8	7.84
3.5	12.25	3.0	9
3.4	11.56	3.5	12.25
3.1	9.61	3.7	13.69
4.0	16	3.3	10.89
3.8	14.44	3.0	9
3.0	9	3.1	9.61
4.1	16.81	2.9	8.41
4.0	16	2.8	7.84
35	124.32	31	96.94

$$\bar{x}_1=\frac{\Sigma x_1}{n_1}=\frac{35}{10}=3.5 \qquad \bar{x}_2=\frac{\Sigma x_2}{n_2}=\frac{31}{10}=3.1$$

$$\Sigma x_1^2-\frac{(\Sigma x_1)^2}{n_1}=124.32-\frac{(35)^2}{10}=124.32-122.5=1.82$$

$$\Sigma x_2^2-\frac{(\Sigma x_2)^2}{n_2}=96.94-\frac{(31)^2}{10}=96.94-96.1=0.84$$

Degrees of freedom, d.f. = $(n_1 + n_2 - 2) = (10 + 10 - 2) = 18$

$$\text{Pooled variance, } s_1^2=\frac{1.82+0.84}{18}=\frac{2.66}{18}=0.147$$

$$\text{Pooled standard deviation, } s_d=\sqrt{s_d^2}=\sqrt{0.147}=0.38$$

$$\therefore t=\frac{|\bar{x}_1-\bar{x}_2|}{s_d\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}=\frac{|3.5-3.1|}{0.38\sqrt{\frac{1}{10}+\frac{1}{10}}}=\frac{0.4}{0.38\sqrt{\frac{1}{5}}}=\frac{0.4}{0.38\times0.45}=\frac{0.4}{0.17}=2.35$$

Conclusion: Calculated t value = 2.35

Scanned with CamScanner

Tabulated t value = 2.10 (p = 0.01)
d.f. = 18

The calculated 't' value for the difference of the two sample means is greater than the tabulated 't' value at 5 % level of probability with 18 d.f. The difference between the two sample means is significant. The null hypothesis stating that the two sample means are not different is rejected at $p < 0.05$. Therefore, gamma-irradiation has a very significant effect on 200 seed weight.

8.1.4.3. Paired t-test

The paired t-test is a statistical test that compares the means and standard deviation of 2 related groups to find whether there is a significant difference between them. It is also known as dependent or correlated t-test. The significant difference happens when the difference between two groups are unlikely to be due to sampling error. A group can be related to similar group of persons or the same items.

A paired t-test are more better than unpaired t-tests as using the similar participants or item eliminates variation among the samples that could be caused by anything other. To study the effect of some action or process, this test has to be used. **For example,** when a stimulus is given to some patients and their blood pressure after and before giving the stimulus is tested, and then with help of paired test it is analysed whether the stimulus is effective or not. Also, the sales of some departmental stores are recorded for definite period after and before some sales advertising procedure and then it is analysed whether the sales advertising procedure is effective in increasing the sales. In similar manner, the performance of a batch of students is recorded before and after training and it is checked if training is effective or not.

Assume that $x_1, x_2, \ldots, x_n$ denotes the sales of a product in n departmental stores for a specific period before sales advertising and $y_1, y_2, \ldots, y_n$ the corresponding sales of the same departmental stores for the same period length after applying certain sales advertising procedure. Then $(x_1, y_2), (x_2, y_2), \ldots, (x_n, y_n)$ are sets of sales data before and after the sales advertising in the n departmental stores.

We apply paired t-test to test the significance of the difference between the sales x and y. Suppose $d_i = x_i - y_i$ or $y_i - x_i$, $i = 1, 2, \ldots, n$.

The differences are $x_i - y_i$ or $y_i - x_i$ so Σd_i is positive. In case, Σd_i is negative we take positive value of d, i.e., |d|.

Often the use of dependent or paired samples will enable us to perform a more precise analysis, because they will allow us to control the extraneous factors. The t-test based on paired observations is defined by the following formula:

$$t = \frac{\bar{d} - 0}{S} \times \sqrt{n} \text{ or } t = \frac{\bar{d}\sqrt{n}}{S}$$

Where, $\bar{d}$ = Mean of the difference,
S = Standard deviation of the differences

The value of S is calculated as follows:

$$S = \sqrt{\frac{\Sigma(d - \bar{d})^2}{n-1}} \text{ or } \sqrt{\frac{\Sigma d^2 - n(\bar{d})^2}{n-1}}$$

It should be noted that t is based on n – 1 degree of freedom.

Procedure

Step 1: Set up the null hypothesis.

Step 2: Then we find the value of standard deviation from the formula:

$$S = \sqrt{\frac{\Sigma(d - \bar{d})^2}{n-1}} \qquad \bar{d} = \text{mean of difference}$$

Step 3: Degree of freedom = n – 1 obtained the table value at the level of significance.

Step 4: Putting this value of standard deviation, find the value of t.

Step 5: If t_{cal} < table value at the level of significance, then null hypothesis is accepted.

If t_{cal} > table value at the level of significance, then null hypothesis is rejected and alternative hypothesis is accepted.

Hypotheses of an Paired t-Test

1) **H_0 (Null Hypothesis):** "There is no significant difference between the means of 2 groups."
2) **H_1 (Alternative Hypothesis):** "There is a significant difference between 2 population means, and this difference is unlikely to be caused by sampling error."

Assumptions of a Paired t-Test

1) The dependent variable is distributed normally
2) The observations are sampled independent
3) The dependent variable is measured on an incremental level like ratios.
4) Independent variables must consist of two matched pairs.

Example 5: An Aquarius health club has been promoting a demanding programme for body conditioning. The club claims that after 1 month, the average participant should be more able to do least eight push-ups in 2 minutes than he or she could do at the beginning. Does the club's claim is right on the random sample of ten programme participants given below? Take the 0.025 level of significance.

Participant	1	2	3	4	5	6	7	8	9	10
Before	38	11	34	25	17	38	12	27	32	29
After	45	24	41	39	30	44	30	39	40	41

Solution: H_0(Null Hypothesis): $\mu_1 - \mu_2 = 8$
H_1(Alternative Hypothesis): $\mu_1 - \mu_2 > 8$ (The difference is more than 8)

Table 8.2: α = 2.5% Level of Significance

Participant	Before	After	Difference = d	(Difference)^2 = (d^2)
1	38	45	+7	49
2	11	24	+13	169
3	34	41	+7	49
4	25	39	+14	196
5	17	30	+13	169
6	38	44	+6	36
7	12	30	+18	324
8	27	39	+12	144
9	32	40	+8	64
10	29	41	+12	144
Total			110	1344

Scanned with CamScanner

$$\bar{d} = \frac{\Sigma d}{n} = \frac{+110}{10} = +11; \quad SE = \sqrt{\frac{n\Sigma d^2 - (\Sigma d)^2}{n\sqrt{n-1}}} = \sqrt{\frac{10\times1344 - (110)^2}{10\times3}} = 1.22$$

One tailed test is used here, as the club needs to prove that the increase is more than 8. An appropriate t-value at 2.5 per cent significance level with 9 d.f is the t-value for two tailed test at 0.05 significance level with 9 d.f. and is equal to 2.262.

The upper limit = 8 + 2.262(1.22) = 10.76

As the mean difference 11 lies outside the acceptance region, hence, null hypothesis H_0 is rejected. This result supports the club's claim.

Example 6: A random sample of 8 companies was taken from companies having at least 15 per cent dividends. Then it has to test whether the statement of dividends has any effect on the market price of a share of a company. A data of share prices of 8 companies are as follows:

	1	2	3	4	5	6	7	8
Market Price 10 Days before Dividends were Declared	70	65	112	58	25	147	95	68
Market Price 10 Days after the Declaration of Dividends	80	85	110	64	32	159	100	70

1) Can we predict that the increase in average market price after declaration of dividends is significant at 5% level of significance? Consider that population is normal distributed.
2) List the all possible inadequate in the methodology accepted to test the validity of the above statement (1).

Solution: This questions is based on paired t-test:

1) H_0: The increase in market price after declaration of dividends is not significant.
   H_1: The increase in market price after declaration of dividends is significant.

   Let us suppose that X and Y indicate the market price 'before' and 'after' declaration of dividends then

   $H_0: \mu_1 - \mu_2$

   [μ_1, μ_2 is the average market price before and after the declaration of the dividend]

   $H_1 = \mu_1 > \mu_2$

   That is,

   $H_0: \mu_1 - \mu_2 = 0$

   $H_1: \mu_2 > \mu_1$

   Assume d = Y – X, where

X	Y	d	d^2
70	80	10	100
65	85	20	400
112	110	–2	4
58	64	6	36
25	32	7	49
147	159	12	144
95	100	5	25
68	70	2	4
		60	762

$$\bar{d} = \frac{60}{8} = 7.5; \quad t = \frac{|\bar{d}|n\sqrt{n-1}}{\sqrt{n\Sigma d^2 - (\Sigma d)^2}} = \frac{7.5\times8\times\sqrt{7}}{\sqrt{8\times762 - (60)^2}} = \frac{158.75}{50} = 3.18$$

The experiential t value is $t = \frac{7.5}{2.3569} = 3.18$ (with 7 degree of freedom).

The value 1.895 is the tabulated value of t at 5% level of significance for 7 degree of freedom.

Tabulated value of t (1.895) < calculated t value (3.18)
Thus, we discard H_0.
Hence, after declaration of dividends, there is significant increase in prices.

2) Inadequacies are as follows:
   i) These have too small Sample size.
   ii) The circumstances prevailing on these dates may be different as the dividends may be declared on different dates.
   iii) Although conclusion is that dividends affect market price positively but there is not certain that the dividend is the only cause behind increase.
   iv) A group of companies that have declared 15% or more dividends are considering only. All the companies which do not follow this group are excluded from the study. Their prices might have increment during the same period.

Example 7: A similar test was given to 12 participants to verify that whether a course in quantitative analysis improved performance. Their scores both before and after the course are shown in table below:

Candidate	1	2	3	4	5	6	7	8	9	10	11	12
Before	44	40	61	52	32	44	70	41	67	72	53	72
After	53	38	69	57	46	39	73	48	73	74	60	78

Suppose these twelve candidates as a sample from a population and test at 5% level of significance in case the course was useful in terms of presentation on the test.

Solution: We apply here the paired t-test.

Candidate	1	2	3	4	5	6	7	8	9	10	11	12
Score before the Course	44	40	61	52	32	44	70	41	67	72	53	72
Score after the Course	53	38	69	57	46	39	73	48	73	74	60	78
Difference (d)	9	–2	8	5	14	–5	3	7	6	2	7	6
d^2	81	4	64	25	96	25	9	49	36	4	49	36

This is the problem of one tailed test at 5% significance and 11d.f., $t_{0.05} = 1.796$ (table value):

1) **Null Hypothesis:** There is no improvement due to course.
2) **Alternate Hypothesis:** Course has better performance.

$\Sigma d = 60$

$$\frac{\Sigma d}{n} = \frac{9-2+8+5+14-5+3+7+6+2+7+6}{12} = \frac{50}{12} = 5$$

$$\Sigma d^2 = 81+4+64+25+196+25+9+49+36+4+49+36 = 578$$

$$t = \frac{|\bar{d}|n\sqrt{n-1}}{\sqrt{n\Sigma d^2 - (\Sigma d)^2}} = \frac{5\times12\times\sqrt{11}}{\sqrt{12\times578 - (60)^2}} = \frac{199}{\sqrt{3336}}$$

$$= \frac{199}{57.758} = 3.445$$

Scanned with CamScanner

As the calculated value of t (3.445) > table value (1.796), Hence Null hypothesis is rejected, i.e., the course has better performance.

Example 8: Before given a drug, the Systolic blood pressures of six hypertensive patients were 183, 179, 165, 190, 175 and 180mm of Hg. After given a particular drug for one week, the blood pressures were recorded as 187, 175, 150, 180, 180 and 170mm of Hg respectively. Determine could such differences arise due to chance?

Solution:

Null Hypothesis: H_0: $\mu_1 = \mu_0$ or $\overline{D} = 0$

Alternative Hypothesis: H_1: $\mu_1 \neq \mu_2$.

Calculation for $\overline{D}$ and estimated variance are shown in table below:

Patients	B.P. before Drug (X_1)	B.P. after Drug (X_2)	Difference D = ($X_1 - X_2$)	D^2
1	183	187	–4	16
2	179	175	4	16
3	165	150	15	225
4	190	180	10	100
5	175	180	–5	25
6	180	170	10	100
			$\Sigma D = 30$	$\Sigma D^2 = 482$

Mean Difference $\overline{D} = \frac{\Sigma D}{n} = \frac{30}{6} = 5$

Estimated variance, $s^2 = \frac{1}{n-1}\left[\Sigma D^2 - \left(\frac{\Sigma D}{\sqrt{n}}\right)^2\right] = \frac{1}{6-1}\left[482 - \frac{(30)^2}{6}\right]$

$= \frac{1}{5}\left[482 - \frac{900}{6}\right] = \frac{1}{5}[482 - 150]$

$\therefore \quad s^2 = \frac{1}{5} \times 332 = 66.4$ or $\quad s = \sqrt{66.4} = 8.15$

Standard Error of $\overline{D} = \frac{s}{\sqrt{n}} = \frac{8.15}{2.45} = 3.33$; $t = \frac{\overline{D}}{\text{S.E. of } \overline{D}}$ Or $t = \frac{5}{3.33} \Rightarrow \therefore t = 1.5$

As the calculated value of $|t| = 1.5 < t_{0.05,5} = 4.03$ (i.e., table value of t and 1% of significance difference with 5 d.f.) Thus, the null hypothesis is accepted and we conclude that there is no significance difference between the averages of before and after observations.

Example 9: In order to specify the result of specific oral contraceptive on weight gain, 9 healthy females were weighed before to the start of its use and again after ending of 3 months period.

Subject No.	Initial Weight (kg)	Weight after 3 Months (kg)
1	48.0	49.2
2	56.4	57.2
3	52.0	56.0
4	60.0	58.0
5	54.0	56.0
6	56.0	57.2
7	48.0	47.2
8	56.0	56.4
9	52.0	52.8

Test and provide sufficient evidence that females experience gain in weight after 3 months of the oral contraceptive use?

Solution:

Subject No.	Initial Weight (kg)	Weight after 3 Months (kg)	Difference D	D^2
1	48.0	49.2	–1.2	1.44
2	56.0	57.2	–0.8	0.64
3	52.0	56.0	–4.0	16.00
4	60.0	58.0	2.0	4.00
5	54.0	56.0	–2.0	4.00
6	56.0	57.2	–1.2	1.44
7	48.0	47.2	0.8	0.64
8	56.0	56.4	–0.4	0.16
9	52.0	52.8	–0.8	0.64
Total	–	–	$\Sigma D = -7.6$	$\Sigma D^2 = 28.96$

Mean difference $\overline{D} = \frac{\Sigma D}{n} = \frac{-7.6}{9} = -0.84$

Estimated Variance, $s^2 = \frac{1}{n-1}\left[\Sigma D^2 - \left(\frac{\Sigma D}{\sqrt{n}}\right)^2\right] = \frac{1}{9-1}\left[28.96 - \frac{57.76}{9}\right]$

$= \frac{1}{8}[28.96 - 6.42] = \frac{1}{8} \times 22.54 \Rightarrow \therefore s^2 = 2.82$ or $s^2 = \sqrt{2.82} = 1.68$

Null hypothesis, H_0: There is no weight gain after 3 months.i.e., $\mu_1 = \mu_2$

Alternate Hypothesis, H_1: There is weight gain after 3 months of the oral contraceptive use. i.e., $\mu_1 \neq \mu_2$ (Two Tailed test).

Standard Error of difference, S.E. $(\overline{D}) = \frac{s}{\sqrt{n}}$

$\therefore$ S.E. $(\overline{D}) = \frac{1.68}{\sqrt{9}} = \frac{1.68}{3} = 0.56$

Test statistic, $t = \frac{\overline{D}}{\text{S.E. of } (\overline{D})} = \frac{0.84}{0.56} \quad \therefore t = -1.5$

Since calculated value $|t| = 1.5 <$ table value of $t_{0.01,8} = 3.3$. This implies that the null hypothesis is accepted, i.e., there is no weight gain after 3 months of the oral contraceptive use.

8.2. ANOVA (ANALYSIS OF VARIANCE)

8.2.1. Introduction

The Analysis of Variance is similar to the regression analysis. It is abbreviated as ANOVA, developed by **R.A. Fisher.** ANOVA is widely used by the researchers to investigate the underlying relationship between a dependent variable and various independent variables.

It is widely used in the areas of business, industries, economics, psychology, sociology, biology etc. The researchers prefer the ANOVA when there are multiple cases of sample. It is a statistical tool which is used to analyse the variations among the data and calculate the inequality among the population means.

Scanned with CamScanner

ANOVA also checks the level of homogeneity among the various groups of samples from different populations. The use of ANOVA by researcher helps them to divide the total variation of the various samples in two groups.

1) The first variation is that which occurs due to chance and
2) The second variation is that which occurs due to any specific cause.

ANOVA helps the researcher, to split the variance for investigative purpose. With the use of ANOVA the researchers can examine 'n' number of factors influencing the dependent variables. In simpler words, Analysis of Variance (ANOVA) is a statistical technique which is used for exploring the variation of a dependent variable with respect to independent variables. In the ANOVA technique when single factor is taken and researcher analyses the difference amongst its various categories which have numerous possible values, then this ANOVA technique is known as **one-way ANOVA**. When analysis is done on two factors at the same time, then this ANOVA is known as **two-way ANOVA**.

8.2.2. Characteristics of ANOVA

1) ANOVA research tool is helpful in the study of analysis of variance.
2) One of the most important characteristics of ANOVA is that it splits the variance into chance variance and cause variance.
3) ANOVA is used in the examination of the various comparative experiments.
4) ANOVA results are free from any type of constant bias and scaling errors.
5) It describes the true and complete characteristics of given populations.
6) It examines the variance of more than two groups of samples.

8.2.3. Assumptions of ANOVA

Analysis of variance is done on the basis of some assumptions. These assumptions are given below:

1) Normally distributed universe used to take the sample for the study.
2) Samples are selected randomly and they are also homogeneous and independent of each other.
3) It is assumed that there is no significant difference amongst the variances of the different universes from which the samples have been drawn.
4) Starts with the null hypothesis i.e., $V_1 = V_2 = V_3 = ...V_n$ (where, V_1, V_2, V_3 $....V_n$ are variance)
5) It is assumed that the critical values of the variance ratio, (F) as estimated by Snedecar for different degrees of freedom at different level of significance, viz.; 5%, or 1%, etc., hold good in testing the significance of the calculated values of F that represents the variance ratio of the samples.

8.2.4. Applications of ANOVA

1) With the help of the ANOVA an organic scientist or researcher can elaborate whether different varieties of fertilizers, seeds or soils differs significantly or not. This helps the researchers to draft the policy accordingly to the outcomes.
2) If the medical researchers or scientists are interested to study the impact of the medicine on different age groups of the human beings, then ANOVA is very useful.
3) The top management of an organisation may examine the performance of their sales managers working on different population. Using ANOVA, management can come to a decision whether the performance or the efficiency of the sales managers differs significantly or not.
4) The statistical tool ANOVA can help the researcher to study whether the purchase intentions of group of people differs significantly or not.

8.2.5. Basic Principle of ANOVA

The underlying principle of the ANOVA is to compare the differences in the different means of population, by studying amount of the variations within the samples, with respect to amount of the variation present between samples of the population. In case of the variations which occur inside given population, generally it is presumed by the researcher that this is because of the random effect in the population. In simpler words, because of the random effect within the population, the value of X_{ij} (a random value) differs with mean of that population.

It has been observed that there exist some influences on the (X_{ij}), which cannot be explainable. On the other hand, the differences which occur between different populations are due to some 'specific factor'. This is the reason why the researchers assume using the ANOVA technique; the samples have been selected from the normal population. Also each population has same degree of the variance. It is assumed by the researcher that there are few variables which can't be controlled but rest of variables are easily controllable. So, researcher assumes that absence of various factors may effects the results regarding the factor(s) to be studied.

While estimating the ANOVA, the researcher needs the estimates the two population variance:

1) Based on within Sample Variance and
2) Between the Sample Variance.

The value of the above two estimates of the population are compared with the F-test. The F-test is calculated as follows:

$$F = \frac{\text{Estimation of the population variance based on the samples variance}}{\text{Estimation of the population variance based on within samples variance}}$$

The above calculated value of the F-test is then compared with the limit of the F-test for a given degree of freedom. If the calculated value of the above equation is more than or equal to F-Limit, then there exists a significant change between the sample means.

8.2.6. ANOVA Techniques

Following are the two ANOVA techniques:

1) Variance which occurs due to one variable, and
2) Variance which occurs due to two variables.

The above stated techniques of the variances are discussed in detail which is as follows:

1) **Variance which occurs due to One Variable:** In this case the change or the variance that occurs in a variable is because of only a single variable. Or we can say that there is only one independent variable and we try to study the impact of that independent variable on the other dependent variable.

   For example, if the researcher wants to study the impact of one variable (e.g., any one amongst types of seed or fertilizer or soil) on the other variable (i.e., amount of harvest of a crop), then this is known as one-way ANOVA or variance due to the one variable. Here type of soil, fertilizer and seeds are three independent variables.

   In this case the researcher will study the impact of one independent variable (from seed, fertilizer or soil) on the dependent variable (harvesting of crop). This type of relationship can be given as:

Scanned with CamScanner

ANOVA also checks the level of homogeneity among the various groups of samples from different populations. The use of ANOVA by researcher helps them to divide the total variation of the various samples in two groups.

1) The first variation is that which occurs due to chance and
2) The second variation is that which occurs due to any specific cause.

ANOVA helps the researcher, to split the variance for investigative purpose. With the use of ANOVA the researchers can examine 'n' number of factors influencing the dependent variables. In simpler words, Analysis of Variance (ANOVA) is a statistical technique which is used for exploring the variation of a dependent variable with respect to independent variables. In the ANOVA technique when single factor is taken and researcher analyses the difference amongst its various categories which have numerous possible values, then this ANOVA technique is known as **one-way ANOVA**. When analysis is done on two factors at the same time, then this ANOVA is known as **two-way ANOVA**.

8.2.2. Characteristics of ANOVA

1) ANOVA research tool is helpful in the study of analysis of variance.
2) One of the most important characteristics of ANOVA is that it splits the variance into chance variance and cause variance.
3) ANOVA is used in the examination of the various comparative experiments.
4) ANOVA results are free from any type of constant bias and scaling errors.
5) It describes the true and complete characteristics of given populations.
6) It examines the variance of more than two groups of samples.

8.2.3. Assumptions of ANOVA

Analysis of variance is done on the basis of some assumptions. These assumptions are given below:

1) Normally distributed universe used to take the sample for the study.
2) Samples are selected randomly and they are also homogeneous and independent of each other.
3) It is assumed that there is no significant difference amongst the variances of the different universes from which the samples have been drawn.
4) Starts with the null hypothesis i.e., $V_1 = V_2 = V_3 = \ldots V_n$ (where, V_1, V_2, V_3 $\ldots V_n$ are variance)
5) It is assumed that the critical values of the variance ratio, (F) as estimated by Snedecar for different degrees of freedom at different level of significance, viz.; 5%, or 1%, etc., hold good in testing the significance of the calculated values of F that represents the variance ratio of the samples.

8.2.4. Applications of ANOVA

1) With the help of the ANOVA an organic scientist or researcher can elaborate whether different varieties of fertilizers, seeds or soils differs significantly or not. This helps the researchers to draft the policy accordingly to the outcomes.
2) If the medical researchers or scientists are interested to study the impact of the medicine on different age groups of the human beings, then ANOVA is very useful.
3) The top management of an organisation may examine the performance of their sales managers working on different population. Using ANOVA, management can come to a decision whether the performance or the efficiency of the sales managers differs significantly or not.
4) The statistical tool ANOVA can help the researcher to study whether the purchase intentions of group of people differs significantly or not.

8.2.5. Basic Principle of ANOVA

The underlying principle of the ANOVA is to compare the differences in the different means of population, by studying amount of the variations within the samples, with respect to amount of the variation present between samples of the population. In case of the variations which occur inside given population, generally it is presumed by the researcher that this is because of the random effect in the population. In simpler words, because of the random effect within the population, the value of Xij (a random value) differs with mean of that population.

It has been observed that there exist some influences on the (Xij), which cannot be explainable. On the other hand, the differences which occur between different populations are due to some 'specific factor'. This is the reason why the researchers assume using the ANOVA technique; the samples have been selected from the normal population. Also each population has same degree of the variance. It is assumed by the researcher that there are few variables which can't be controlled but rest of variables are easily controllable. So, researcher assumes that absence of various factors may effects the results regarding the factor(s) to be studied.

While estimating the ANOVA, the researcher needs the estimates the two population variance:

1) Based on within Sample Variance and
2) Between the Sample Variance.

The value of the above two estimates of the population are compared with the F-test. The F-test is calculated as follows:

$$F = \frac{\text{Estimation of the population variance based on the samples variance}}{\text{Estimation of the population variance based on within samples variance}}$$

The above calculated value of the F-test is then compared with the limit of the F-test for a given degree of freedom. If the calculated value of the above equation is more than or equal to F-Limit, then there exists a significant change between the sample means.

8.2.6. ANOVA Techniques

Following are the two ANOVA techniques:

1) Variance which occurs due to one variable, and
2) Variance which occurs due to two variables.

The above stated techniques of the variances are discussed in detail which is as follows:

1) **Variance which occurs due to One Variable:** In this case the change or the variance that occurs in a variable is because of only a single variable. Or we can say that there is only one independent variable and we try to study the impact of that independent variable on the other dependent variable.

   For example, if the researcher wants to study the impact of one variable (e.g., any one amongst types of seed or fertilizer or soil) on the other variable (i.e., amount of harvest of a crop), then this is known as one-way ANOVA or variance due to the one variable. Here type of soil, fertilizer and seeds are three independent variables.

   In this case the researcher will study the impact of one independent variable (from seed, fertilizer or soil) on the dependent variable (harvesting of crop). This type of relationship can be given as:

Scanned with CamScanner

A (In dependent variable)	B (dependent variable)
11	14
12	16
17	11
16	10
14	15

2) **Variance which occurs due to Two Variables:** In this case, the researcher wants to study the impact of multiple independent variables on a single dependent variable. Here, the impact of the changes or the variation in two or more independent variables is studied in different condition and also the values corresponding dependent variable is also observed. **For example,** if a researcher wants to study the impact of two independent variables, say season and effort of the salesman, on the total sale of corresponding salesman. This type of situation is best example of the variation due to the impact of two variables. The above example can be best described with the help of below stated format.

Samples of Sales in '000 ₹

Seasons	Salesmen				Total
	P	Q	R	S	
Summer	29	28	26	25	108
Monsoon	34	33	34	19	120
Winter	25	28	27	28	108
Total	88	89	87	72	336

On the basis of the above classification there are two ANOVA techniques:
1) Analysis of the variance for one way classification (One way ANOVA)
2) Analysis of the variance for two way classification (Two way ANOVA)

8.2.7. One Way ANOVA

In the concept of the one way ANOVA, there is an analysis of variance of a continuous dependent variable with a single classification variable (Independent Variable). One-way analysis of variance (ANOVA) always tests equality of the population means when there is classification only by one variable. After calculations we check whether there exists any difference within the factor or not.

Process for Calculation of One Way ANOVA

1) Calculate the means of all sample individually, that find:

$\bar{X}_1, \bar{X}_2, \bar{X}_3, \ldots\ldots, \bar{X}_k$

Here, k = Samples Observation

2) Calculate the mean of the sample:

$$\bar{\bar{X}} = \frac{\bar{X}_1 + \bar{X}_2 + \bar{X}_3 + \ldots\ldots + \bar{X}_k}{\text{No. of Samples}(k)}$$

3) Then calculate the deviation of the 'sample means' from the 'means of the sample means'. Then each deviation is squared and multiplied by the total number of the items in corresponding sample. Each item is then added up and the output so obtained is called as **Sum of Squares (SS).**

SSB = SS between

$$= n_1\left(\bar{X}_1 - \bar{\bar{X}}\right)^2 + n_2\left(\bar{X}_2 - \bar{\bar{X}}\right)^2 + \ldots\ldots + n_k\left(\bar{X}_k - \bar{\bar{X}}\right)^2$$

4) Then we need to divide the 'output is obtained from third step by 'degrees of the freedom'. Then, we get the mean square between samples (MS). It is calculated as follows:

$$\text{MSB} = \text{MS between} = \frac{\text{SSB}}{(k-1)}$$

Here the MSB is the mean square between the samples, & (k–1) denotes degree of freedom (d.f.).

5) The researcher gets the value of the deviations for all of the sample items from their means. Then calculate the square of all the deviations and sum it up. The value so obtained after addition, is called **as sum of the squares** of the variance within sample, abbreviated as (SS within). It can be written as:

$$\text{SSW} = \text{SS within} = \sum\left(X_{1_j} - \bar{X}_1\right)^2 + \sum\left(X_{2_j} - \bar{X}_2\right)^2 + \ldots\ldots\ldots + \sum\left(X_{n_j} - \bar{X}_k\right)^2$$

6) Then divide the 'output of fifth step' by 'degree of the freedom' within the samples, in order to get the 'mean square' also abbreviated as MS (mean square or the mean square within sample). It is written as:

$$\text{MSW} = \text{MS within} = \frac{\text{SSW}}{(n-k)}$$

Where,
(n – k) = represents degrees of freedom within samples,
n = total number of items in all the samples, i.e., $n_1 + n_2 + \ldots + n_k$
k = number of samples.

7) To cross check, the researcher can calculate the sum of the squares of the deviation for the total variance by adding up the squares of all the deviations. This can be done when the deviations of the each item from each individual sample is determined using 'mean of sample means'. This is represented as:

$$\text{SST} = \text{SS for total variance} = \sum\left(X_{ij} - \bar{\bar{X}}\right)^2;$$

$i = 1,2,3,\ldots$; $j = 1,2,3,\ldots\ldots$

This value of SST is equal to the total of the results of the step three and five. In each case the degree of the freedom is calculated by subtracting 1 from the total number of items in each sample i.e. (n–1). The degree of the freedom in the case of within and between the samples should add to degree of the freedom for the complete variance, this can be represented as: $(n-1) = (k-1) + (n-k)$

This fact describes the additive property of the ANOVA.

8) Finally, the F-ratio can be represented as:

$$\text{F-ratio} = \frac{\text{MSB}}{\text{MSW}}$$

The researcher uses this ratio to know if the differences among various samples mean is just because of the sampling fluctuation or it has some significance and cannot be ignored. The researchers consult the table and put the F value for given 'degrees of the freedom' at different significance levels. If the calculated value of F is less than table value of F, then the differences are insignificant and can be considered that it is occurring just by chance and the researcher accepts the Null Hypothesis. If the calculated value of F is more than or equal to the table value of F then the results are significant and the researcher accepts the alternate hypothesis.

Scanned with CamScanner

Example 10: Four machines A, B, C and D are used to produce a certain kind of cotton fabrics. Samples of size 4 with each unit as 100 square metres are selected from the outputs of the machines at random, and the number of flaws in each 100 square metres are counted, with the following result:

A	B	C	D
11	7	12	16
9	9	11	20
14	10	18	23
16	4	9	21

Do you think that there is a significant difference in the performance of the four machines?

Solution. Let us take the null hypothesis that the machines do not differ significantly in performance, i.e., $H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$

X_1	X_1^2	X_2	X_2^2	X_3	X_3^2	X_4	X_4^2
11	121	7	49	12	144	16	256
9	81	9	81	11	121	20	400
14	196	10	100	18	324	23	529
16	256	4	16	9	81	21	441
$\Sigma X_1 = 50$	$\Sigma X_1^2 = 654$	$\Sigma X_2 = 28$	$\Sigma X_2^2 = 246$	$\Sigma X_3 = 50$	$\Sigma X_3^2 = 670$	$\Sigma X_4 = 80$	$\Sigma X_4^2 = 1626$

$T = \Sigma X_1 + \Sigma X_2 + \Sigma X_3 + \Sigma X_4 = 50 + 28 + 50 + 80 = 208$

$$\text{Correction Factor} = \frac{T^2}{N} = \frac{(208)^2}{16} = \frac{44521}{16} = 2704$$

$$SST = \Sigma X_1^2 + \Sigma X_2^2 + \Sigma X_3^2 + \Sigma X_4^2 - \frac{T^2}{N} = 654 + 246 + 670 + 1626 - 2704 = 3196 - 2704 = 492$$

$$SSB = \frac{(\Sigma X_1)^2}{n_1} + \frac{(\Sigma X_2)^2}{n_2} + \frac{(\Sigma X_3)^2}{n_3} + \frac{(\Sigma X_4)^2}{n_4} - \frac{T^2}{N}$$

$$= \frac{(50)^2}{4} + \frac{(28)^2}{4} + \frac{(50)^2}{4} + \frac{(80)^2}{4} - 2704 = 625 + 196 + 625 + 1600 - 2704 = 342$$

$SSW = SST - SSB = 492 - 342 = 150$

ANOVA Table			
Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square
Between samples	342	3	114
Within samples	150	12	12.5
Total	492	15	

$$F_{(3,12)} = \frac{114}{12.5} = 9.12$$

The table value for $F_{(3,12)}$ at 1% level of significance is 5.95. The calculated value of F is greater than the table value. Hence, we reject the null hypothesis and conclude that there is a significant difference in the performance of the four machines.

Example 11: A common test was performed on a number of patients taken from the four hospitals randomly to measure the significance of possible difference in performance of an antibiotic after its administration to TB patients. The data are given in table below. Give an analysis of variance of data.

Hospital			
A	B	C	D
8	12	18	13
10	11	12	9
12	9	16	12
8	14	6	16
7	4	8	15

Solution: Null Hypothesis, H_0: There is no significant difference in the average effects of antibiotics on TB patients.

Alternate Hypothesis, H_1: There is significant difference in the average effects of antibiotics on TB patients.

A	B	C	D
8	12	18	13
10	11	12	9
12	9	16	12
8	14	6	16
7	4	8	15
$T_1 = 45$	$T_2 = 50$	$T_3 = 60$	$T_4 = 65$

Grand Total $(G) = T_1 + T_2 + T_3 + T_4 = 45 + 50 + 60 + 65 = 220$

$$\text{Correction factor (C.F.)} = \frac{G^2}{n} = \frac{(220)^2}{20} = 2420$$

Total addition of squares $(TSS) = \Sigma\Sigma y_{ij}^2 - C.F.$

$= (8)^2 + (10)^2 + (12)^2 + \ldots\ldots + (16)^2 + (15)^2 - C.F.$

$= 64 + 100 + 144 + \ldots\ldots + 256 + 225 - 2420 = 2678 - 2420 = 258$

Sum of squares among the samples (SST)

$$= \sum_i \frac{T_i^2}{n_i} - C.F. = \frac{(45)^2}{5} + \frac{(50)^2}{5} + \frac{(60)^2}{5} + \frac{(65)^2}{5} - 2420 = 2470 - 2420 = 50$$

Sum of squares within the samples (Error)

$SSE = TSS - SST = 258 - 50 = 208$

The ANOVA table is shown below:

Table: ANOVA Table

Source of Variation	d. f.	Sum of Squares	Mean Sum of Squares	Variance Ratio F
Between samples	4 – 1 = 3	50	$\frac{50}{3} = 16.7$	$F = \frac{16.7}{13} = 1.28$
Within samples	20 – 4 = 16	208	$\frac{208}{16} = 13.0$	
Total	20 – 1 = 19			

Table value of $F_{0.05}$ for (3, 16) degree of freedom (d.f.) is 3.24. Hence the calculated value of F < tabulated value of F, null hypothesis is accepted at 5% level of significance. Thus, the average effect of antibiotics on TB patients in various hospitals of the city is not significant.

8.2.8. Two Way ANOVA

The two way Analysis of Variance (ANOVA) technique is widely used by the researcher when data is categorised in to two factors.

Scanned with CamScanner

The two way ANOVA can be better explained by the following **examples**:

1) A business manager may be interested in analysing the sales data of their sales manager on the basis of the efforts done by them or the area in which they are working.
2) An agricultural researcher or scientist may analyse the rice yield on the basis of the type of seed or fertilizer which is used by the farmers.
3) In a manufacturing unit, the project manager may analyse the output of a machine on the basis of the experience of the worker and the quality of the raw material.

In all the cases the researcher takes the observation at regular interval and makes a note of it. Sometimes, the researcher needs to take some repeated values also.

The ANOVA technique is little different in case of **repeated measurements** where we also compute the interaction variation. Two types of ANOVA techniques are:

1) **ANOVA Technique in the Context of Two way Design when Repeated Values are not there:** In the case when repeated values are missing, researcher cannot compute sum of the squares directly within the sample. So, the residual or the error variation is calculated by subtracting the sum of variance between varieties of a treatment and variance between varieties of other treatment from the total sum of squares for total variance.

Steps

i) The researcher needs to use a coding device, to make the work simpler.
ii) After this, researcher has to take total of all the individual values (or the coded values) of the entire sample, and name it as T.
iii) Then apply correction factor stated below:

$$\text{Correction factor} = \frac{(T)^2}{n}$$

iv) Calculate the square value of all the items (or square of the coded values) taken individually and then sum it up. Then subtract the correction factor from the summation to get the square of deviations for the total variance:

$$\text{Sum of squares of Deviations for total variance or total SS} = \sum X_{ij}^2 - \frac{(T)^2}{n}$$

v) After the above step, calculate the total of all the different columns, and then calculate the square of sum of every column. Then divide the square value of the each column by the corresponding number of items in that column and again get the sum of that particular column. Now, subtract the above calculated correction factor from this total. So the output received is the sum of the squares of deviation for the variance between the columns (SS between columns).

vi) In this step, the researcher calculates the total rows and then finds out the 'square of sum of every row'. This squared sum is then divided by the number of items of that specific row to obtain the sum of that entire row. Now, correction factor is subtracted from the calculated sum and the researcher gets 'sum of the square of the deviation for variance between the columns (SS between columns)'.

vii) Here by subtracting the sum of 5th and 6th step from 4th step researcher can calculate the sum of the squares of the deviation for the error or residual variance.

Total SS – (SS between columns + SS between rows) = SS for residual or error variance

Degrees of freedom (d.f.) can be worked out as under:

d.f. for Total Variance = $(c, r - 1)$
d.f. for Variance between Columns = $(c - 1)$
d.f. for Variance between Rows = $(r - 1)$
d.f. for Residual Variance = $(c - 1)(r - 1)$

Where, c = Number of Columns
r = Number of Rows

viii) ANOVA table can set up as:

Analysis of Variance Table for Two Way ANOVA

Source of Variation	Sum of Squares (SS)	Degrees of the Freedom (d.f.)	Mean Square Value(MS)	F-Ratio Value
Between the columns Treatment	$\sum \frac{(T_j)^2}{n_j} - \frac{(T)^2}{n}$	$(c - 1)$	$\frac{SSB(\text{columns})}{(c-1)}$	$\frac{MSB(\text{columns})}{MSR}$
Between the rows treatment	$\sum \frac{(T_i)^2}{n_i} - \frac{(T)^2}{n}$	$(r - 1)$	$\frac{SSB(\text{columns})}{(r-1)}$	$\frac{MSB(\text{rows})}{MSR}$
Residual or error	Total SS – (SS between the columns + SS between the rows)	$(c - 1)(r - 1)$	$\frac{SSR}{(c-1)(r-1)}$	
Total	$\sum X_{ij}^2 - \frac{(T)^2}{n}$	$(c\,r - 1)$		

Where,

c = Number of the Columns
r = Number of the Rows
SSR (SS residual) = Total SS – (SS between the columns + SS between the rows)

The basis for F-ratio is provided by the Mean Square Residual (MSR) or residual variance. It is concerned with the variations that are between the column and the row calculation. The fluctuations in the sampling always cause the Mean Square residual. The two F-ratios are then compared with the related table values, for a given degree of the freedom at specified significance level. If the calculated value for the F-ratio is greater than or equal to the table value, in that case, difference of column means is treated significant. In the same way the F-ratio related to the variations between the rows can be calculated.

2) **ANOVA Technique in the Context of Two Way Design when Repeated Values are there:** The researcher gets the separate value of the individual smallest change or variation, in the case of the two way ANOVA, when there are some repeated values. The degree of freedom and sum of the squares are calculated for the variance within the sample same as in case of one way ANOVA. The total sum of the square (SS), which are between the columns and the rows, can be calculated as stated above. Later, the researcher finds out the 'left over sums of the square' and also the 'left over degree of freedom' which is known as the 'interaction variation' (Interaction variation is the value of the relationship present between two different classifications). After all the calculations have been done, ANOVA table can be set up for observing various inferences.

Scanned with CamScanner

Example 12: 5 doctors each test 5 treatments for a certain disease and observe the number of days each patient takes to recover. The results are given in below table. Discuss the difference between

1) The doctors, and
2) The treatment.

Use $\alpha = 0.05$.

Doctors	Treatments				
	1	2	3	4	5
1	10	14	23	18	20
2	11	15	24	17	21
3	9	12	20	16	19
4	8	13	17	17	20
5	12	15	19	15	22

Solution: We subtracted 15 from the given values and work out with the new values of x_{ij}.

Doctors	Treatments					
	1	2	3	4	5	Total
1	−5	−1	8	3	5	10
2	−4	0	9	2	6	13
3	−6	−3	5	1	4	1
4	−7	−2	2	2	5	0
5	−3	0	4	0	7	8
Total	−25	−6	28	8	27	32

Correction Factor $= \frac{T^2}{N} = \frac{(32)^2}{25} = 40.96$ (Where number of observations or N is 25)

Sum of Squares between Columns (Treatments)

This will be obtained by squaring the treatments totals, dividing each total by the number of observations included in it, adding these figures and then subtracting the correction factor from them. Thus, sum of squares between salesmen is:

$$= \frac{(-25)^2}{5} + \frac{(-6)^2}{5} + \frac{(28)^2}{5} + \frac{(8)^2}{5} + \frac{(27)^2}{5} - \frac{T^2}{N}$$

$$= 125 + 7.2 + 156.8 + 12.8 + 145.8 - 40.96 = 406.64$$

Degrees of Freedom $(v) = (5 - 1) = 4$

Sum of Squares between Rows (Doctors)

This will be obtained by dividing the squares of the doctors totals by the number of observations that make-up each total, adding all such figures and subtracting these from the correction factor. Thus, sum of squares between doctors is:

$$= \frac{(10)^2}{5} + \frac{(13)^2}{5} + \frac{(1)^2}{5} + \frac{(0)^2}{5} + \frac{(8)^2}{5} - \frac{T^2}{N} = 66.8 - 40.96 = 25.84$$

Degrees of Freedom $(v) = (5 - 1) = 4$

Total Sum of Squares

This will be obtained by adding the squares of all the observations in the table and subtracting the correction factor therefrom. Thus, total sum of squares is:

$$= (-5)^2 + (-4)^2 + (-6)^2 + (-7)^2 + (-3)^2 + (-1)^2 + (0)^2 + (-3)^2 + (-2)^2 + (0)^2 + (8)^2 + (9)^2 + (5)^2 + (2)^2 + (4)^2 + (3)^2 + (2)^2 + (1)^2 + (2)^2 + (0)^2 + (5)^2 + (6)^2 + (4)^2 + (5)^2 + (7)^2 - \frac{T^2}{N}$$

$$= 25 + 16 + 36 + 49 + 9 + 1 + 0 + 9 + 4 + 0 + 64 + 81 + 25 + 4 + 16 + 9 + 4 + 1 + 4 + 0 + 25 + 36 + 16 + 25 + 49 - 40.96 = 508 - 40.96 = 467.04$$

$v = (25 - 1) = 24$

The above information is presented in the following table:

ANOVA Table

Source of Variation	Sum of Squares	d.f.	Mean Square	F-Ratio
Between columns (Treatments)	406.64	4	101.66	$\frac{101.66}{2.16} = 47.06$
Between rows (Doctors)	25.84	4	6.46	$\frac{6.46}{2.16} = 2.99$
Residual	34.56	16	2.16	
Total	467.04	24		

The table value of F for 4 and 16 degrees of freedom at 5% level of significance is 3.01.

1) Table value of F is greater than the calculated value for doctors. Hence there is not any difference between doctors.
2) Table value of F is less than the calculated value for doctors. Hence there is highly significant difference between treatments.

Example 13: There are four different drugs which have been developed for the cure of a certain disease. These drugs are experimented on patients of three different hospitals. The numbers of cases of recovery from the disease per 100 people are given in table below. Analyse the variance and interpret results.

Hospital	Drugs			
	A	B	C	D
H_1	24	20	24	17
H_2	20	25	30	9
H_3	13	18	31	13

Solution: Null Hypothesis: H_0: There is no result on the drugs in the cure of disease of patients.

Alternate Hypothesis: H_1: There is a result on the drugs in the cure of disease of patients.

Hospital	Drugs				Total
	A	B	C	D	
H_1	24	20	24	17	85
H_2	20	25	30	9	84
H_3	13	18	31	13	75
Total	57	63	85	39	244

Grand total (G) = 244

Correction factor (CF) $= \frac{G^2}{N} = \frac{(244)^2}{12} = \frac{59536}{12} = 4961.33$

Total sum of squares (TSS) $= \sum\sum y_{ij}^2 - C.F. = (24)^2 + (20)^2 + \ldots\ldots + (13)^2 - C.F.$

$= 5470 - 4961.33 = 508.67$

Sum of squares due to rows (SSR) $= \frac{(85)^2}{4} + \frac{(84)^2}{4} + \frac{(75)^2}{4} - C.F.$

$= 1806.25 + 1764 + 1406.25 - 4961.38 = 4976.50 - 4961.33 = 15.17$

Scanned with CamScanner

Sum of squares due to columns (SSC) $= \frac{(57)^2}{3} + \frac{(63)^2}{3} + \frac{(85)^2}{3} + \frac{(39)^2}{3} - C.F.$

$= 1083 + 1323 + 2408.33 + 507 - 4961.33 = 5321.33 - 4961.33 = 360$

Sum of squares due to error (SSE) = TSS – SSR – SSC

$= 508.67 - 15.17 - 360 = 508.67 - 375.17 = 133.5.$

Table: Analysis of Variance Table (ANOVA)

Source of Variation	d.f.	Sum of Squares (SS)	Mean Sum of Squares (MSS)	Variance Ratio F
Between hospitals (rows)	3 – 1 = 2	15.17	$\frac{15.17}{2} = 7.58$	$F_{(2,6)} = \frac{7.58}{22.25} = 0.34$
Between drugs (columns)	4 – 1 = 3	360	$\frac{360}{3} = 120$	$F_{(3,6)} = \frac{120}{22.25} = 5.39$
Error	(3 – 1) (4 – 1) = 6	133.5	$\frac{133.5}{6} = 22.25$	
Total	12 – 1 = 11	508.67		

1) **Tabulated Values:** At 5% level of significance, $F_{(2,6)} = 5.14$ and $F_{(3,6)} = 4.76$.
2) **F for between Drugs:** The calculated value of F > tabulated value of F, thus, null hypothesis is rejected. This shows that the drugs have significant means.
3) **F for between Hospitals:** The calculated value of F < tabulated value of F, hence the null hypothesis is accepted. In other words, the hospitals have no significant means.

8.3. FISHER'S LEAST SIGNIFICANT DIFFERENCE (LSD) TESTS

This method is given by **R.A. Fisher** and considered as the modest approach to a posteriori comparisons. This test for comparing all possible pairs of treatment means is simple t-test. Though, the LSD procedure outcomes in more significant differences than would be expected according to the level. The LSD test varies from the random use of various t-tests in that one proceeds only if the F test in the ANOVA is significant, and the pooled variance is used as the variance estimate in the t-test method.

According to many researchers for comparing the various means for easy holding and then to the accuracy of the access to the correct results, LSD is considered as the best process. This is also known as **Protected Least Significant Difference**, because the **Fisher** does not recommend using the test unless the F-cal for treatments is significant.

LSD test which is used to compare differences between two means is called as original T-test. It is used to test, for every pair of treatments (i, j) in 1,... k:

H_0: $\mu_i = \mu_j$ against H_1: $\mu_i \neq \mu_j$.

According to **Fisher**, the test-statistic to find out whether $\mu_i \neq \mu_j$ is a mild variation of the one used to test whether $\mu_1 \neq \mu_2$ when variances were equal. The test-statistic is represented as:

$$t = \frac{\overline{X}_i - \overline{X}_j}{\sqrt{MSE\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}}$$

The Mean Squared Error (MSE) is given by:

$$MSE = \frac{SSE}{n-k}, \text{ where } SSE = (n_1-1)\cdot s_1^2 + (n_2-1)\cdot s_2^2 + \ldots + (n_k-1)\cdot s_k^2$$

MSE is a valid "pooled variance" estimator, under the maintained assumption that all variances are the same.

Under null hypothesis H_0, the approximate distribution of t is a Student-t with n-k degrees of freedom.

Hence, we should discard H_0 in favour of H_1 in case:

$$|t| > t_{\alpha/2,n-k}$$

Therefore, the least significant difference (LSD) can be defined as:

$$LSD = t_{\alpha/2,n-k}\sqrt{MSE\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$$

Now we conclude that $\mu_i \neq \mu_j$ in case:

$$\left|\overline{X}_i - \overline{X}_j\right| > LSD$$

Example 14: The Investment behaviour between different age groups is given below. Consider that we need to investigate whether the mean proportion of financial assets that is invested in the stock market differs across age group. We consider the following:

1) Young (less than 35)
2) Early middle age (35 to 49)
3) Late middle age (50 to 65)
4) Senior (older than 65)

These are the four different "treatments".

For i = 1,2,3,4, let μ_i = Mean proportion of invested in stock market for individuals in the i^{th} age categories.

Following are total six pairs for the pairwise comparisons:

1) μ_1 and μ_2
2) μ_1 and μ_3
3) μ_1 and μ_4
4) μ_2 and μ_3
5) μ_2 and μ_4
6) μ_3 and μ_4

Here we consider 5% significance level.

For each pair, the LSD critical values are illustrated in table below:

Pair	LSD Critical Value	$\left\|\overline{X}_i - \overline{X}_j\right\|$	Reject H_0: $\mu_i = \mu_j$ for H_1: H_1: $\mu_i \neq \mu_j$?
μ_1 and μ_2	5.8	8.1	Yes
μ_1 and μ_3	6.2	6.7	Yes
μ_1 and μ_4	7.1	7.4	Yes
μ_2 and μ_3	5.6	1.3	No
μ_2 and μ_4	6.5	0.6	No
μ_3 and μ_4	6.9		

Scanned with CamScanner

For all occurrences shown in table above, we get:

$$\text{LSD: } t_{n_1+n_2-k}\sqrt{MSE\left(\frac{1}{n_i}+\frac{1}{n_j}\right)} = t_{\alpha=0.05}\sqrt{447.2\frac{1}{n_i}+\frac{1}{n_j}} = 1.96\sqrt{447.2\left(\frac{1}{n_i}+\frac{1}{n_j}\right)}$$

8.4. EXERCISE

8.4.1. Very Short Answer Type Questions

1) What are parametric tests?
2) Give a difference between assumption parametric test and 't' test.
3) List the type of t –test.
4) What is the pooled t-test?
5) Write the characteristics of ANOVA.
6) Give an application of ANOVA.

8.4.2. Short Answer Type Questions

1) Write the properties of t – distribution.
2) Explain the paired t-test.
3) What is ANOVA (analysis of variance)? And write the assumptions of ANOVA.
4) Define the one way ANOVA and two ways ANOVA.
5) What do you mean fisher's least significant difference (LSD) tests?

8.4.3. Long Answer Type Questions

1) Explain the t-TEST (student's 't' distribution) and give an application. graph of 't' distribution.
2) Samples of two types of electric light bulbs were tested for length of life and following data were obtained:

	Type I	Type II
Sample Number	$n_1 = 8$	$n_2 = 7$
Sample Mean	$\bar{X}_1 =$ 1234 hours	$\bar{X}_2 =$ 1036 hours
Sample S.D.'s	$s_1 =$ 36 hours	$s_2 =$ 40 hours

Is the difference in the means sufficient to warrant that type I is superior to type II regarding length of life?

[Ans: Hypothesis is rejected. $t_{0.05}(13) = 2.16$]

3) Two horses A and B were tested according to time (in seconds) to run a particular race with the following results:

Horse A	28	30	32	33	33	29	34
Horse B	29	30	30	24	27	29	

Test whether you can discriminate between the two horses.
(Given $t_{0.05}(11) = 2.2$)

[Ans: Hypothesis is rejected at 5%]

4) Two independent samples of 8 and 7 items respectively had the following values of the variable (weight in ounces).

Sample-I	9	11	13	11	15	9	12	14
Sample-II	10	12	10	14	9	8	10	

Is the difference between the means of the sample significant?
(Given $t_{0.05}(13) = 2.16$)

[Ans: No]

5) 10 persons were appointed in an electrical position in an office. Their performance was noted by giving a test and means recorded out of 50. They were given 6 months training and again they were given a test and marks were recorded out of 50.

Employees	1	2	3	4	5	6	7	8	9	10
Before Training	25	20	35	15	42	20	20	44	35	48
After Training	25	20	34	13	43	40	29	41	36	46

By applying t-test can it be concluded that the employees have benefited by training ($t_{0.05}(9) = 2.262$).

[Ans: No]

6) There are three main brands of a certain powder. A set of its 120 sales is examined and found to be allocated among four groups (A, B, C, D) and brands (I, II, and III) as shown below.

Brands	Groups			
	A	B	C	D
I	0	4	8	15
II	5	8	13	6
III	18	19	11	13

Is there any significant difference in brands preference? Answer at 5 per cent level using one way ANOVA.

[Ans: No significant difference in brands]

Scanned with CamScanner