Uttar Pradesh Rajarshi

Tandon Open University

Masters of Commerce

M.Com-301

BUSINESS STATISTICS

CONTENTS

.Com -301 BUSINESS STATISTICS

© UPRTOU, Prayagraj. 2024 **First Edition:** July 2024

ISBN:

UNIT 1 INTRODUCTION TO STATISTICS

UNIT STRUCTURE

1.0 OBJECTIVES:

After going through this unit you will be able to understand the

- Meaning of statistics
- Characteristics, functions, division and merits of statistics
- Uses, importance, scope and limitation of statistics

1.1 INTRODUCTION

Statistics is a mathematical science pertaining to the collection, analysis, interpretation or explanation and presentation of data. It provides tools for predicting and forecasting the economic activities. It is useful for an academician, government, business etc. On the basis of various definitions provided by economists, statistics has been broadly defined in two senses: first is in plural sense and the second is in singular sense. In plural sense, statistics refers to numerical facts and figures collected in a systematic manner with a specific purpose in any field of study. In this sense, statistics is also aggregates of facts expressed in numerical form. The characteristics about statistical facts are:

Aggregate of facts

- Numerically expressed
- Data affected by multiplicity of causes

- Enumerated according to reasonable standard of accuracy
- Collected in systematic accuracy
- Collected for pre-determined purpose and
- Placed in relation to other

In singular sense, statistics refers to a science which comprises methods that are used in the collection, analysis, interpretation and presentation of numerical data. These methods are used to draw conclusion about the population parameters. The stages of statistical analysis are:

Stage 1: Collection of data
Stage 2: Organisation of data
Stage 3: Presentation of data
Stage 4: Analysis of data and
Stage 5: Interpretation of data

Statistics helps in business forecasting, decision making, quality control, search of new ventures, study of market, study of business cycles, useful for planning, useful for finding averages, useful for bankers, brokers, insurance, etc.

1.2 STATISTICS MEANING

The term 'statistics' has been derived from the Latin word 'status' Italian word 'statista' or German word 'statistik'.

All these words mean 'Political state'. In ancient days, the states were required to collect statistical data mainly for the number of young men so that they can be recruited in the Army.

Also to calculate the total amount of land revenue that can be collected.

1.3 STATISTICS DEFINITION

Statistics has been defined in different ways by different authors.

- *Statistics are numerical statements of facts in any department of enquiry placed in relation to each other.*

  Bowley

- *By statistics, we mean quantitative data affected to a marked extend by multiplicity of causes*

  Yule and Kendall

- *By statistics, we mean aggregate of facts affected to a marked extent by multiplicity of causes, numerically expressed, enumerated or estimated according to reasonable standards of accuracy, collected in a systematic manner for a predetermined purpose and placed in relation to each other.*

  Horace Secrist

- *Statistics may be defined as the collection, presentation, analysis and interpretation of numerical data.*

  Croxton and Cowden

1.4 CHARACTERISTICS OF STATISTICS

Statistics is characterized by:

1. **Aggregate of Facts**: Single and non-connected facts or figures are not statistics, rather when the facts are aggregates, they are said to be statistics, as they can be compared
2. **Affected to a substantial extent by a variety of reasons**: This means that statistics are influenced to a substantial extent by a number of factors that operate together. **For example**, The statistics of rice production is based on various factors like a method of cultivation, climatic conditions, seeds, fertilizers and manures, etc.
3. **Numerical expression**: Statistics are expressed in terms of numbers. Therefore, qualitative expressions such as happy, sad, right, wrong, good, or bad do not amount to statistics. **For example**: 'Production of ABC ltd. has risen' is not statistics, but 'Production of ABC ltd. has risen from 92000 units in 2020 to 110000 units in 2021' is statistics.
4. **Enumerated and Estimated as per reasonable standard of accuracy**: Reasonable accuracy needs to be there in the statistical data, as it acts as a basis for the field of statistical enquiry. This is because, if the scope of the inquiry is narrow, then by using the method of actual counting, the data can be collected, whereas if the scope of inquiry is wide then the data collection will be based on estimate and estimates can be inaccurate.
5. **Data collection is carried out in a systematic manner**: The collection of statistics should be performed in a systematic as well as planned manner, because in the absence of any system, the data collected can be unreliable and inaccurate, which may also lead to misleading conclusions. Further, the purpose for its collection needs to be stated beforehand to keep its usefulness intact.
6. **Data must be placed in relation to one another**: Data collection is performed for the purpose of comparison and so the basis must be homogeneous. Because when the basis of two units is heterogeneous, the comparison is not possible.

1.5 FUNCTIONS OF STATISTICS

Statistics performs the following functions:

- **Reduces complexities**: Using statistical methods, voluminous data can be presented in a way that it can be easily understood. Hence, it reduces the complexity to understand a vast amount of data, to simplify its meaning.
- **Expresses facts in numbers**: An important function of statistics is that it can transform facts into numbers, which is easy to understand by anyone.
- **Presentation of data in condensed form**: Data collected is usually in raw form, which is complex and unorganized. Hence, it requires to be presented in a simple form so as to reach a final conclusion. With the help of statistics, a large amount of data can be presented in condensed form.
- **Increases the individual knowledge and experience**: As the presentation of data is simple, it enhances the knowledge and experience of people, by making it simple and easy to understand, without having knowledge of each and every field.
- **Different phenomena are compared**: Statistics helps in making a comparison of data and measuring the relationship between them. **For example**: Suppose a researcher wants to

measure the level of production of soybean in two states, then he/she would use statistics.

- **Helpful in the formulation of policies**: Plans and policies are developed beforehand in an organization. And statistics plays a very crucial role in determining the future trends, so as to frame them, by providing the required information.
- **Helpful in prediction and forecasting**: The knowledge of statistics is not just helpful in estimating the present but it also helps in forecasting the future

1.6 DIVISIONS OF STATISTICS

The different types or branches of statistics are discussed hereunder:

1. **Descriptive Statistics**: It involves describing and summarizing the sets of numerical data with the help of pictures and statistical quantities. Techniques used may include averages dispersion, skewness, time series, etc.
2. **Inferential Statistics**: It encompasses those methods that are helpful in drawing conclusion and inferences with respect to parameters of population, based on estimates which are drawn from samples. Chi-square, F-test, t-test, etc techniques are used.
3. **Applied Statistics**: Those methods and techniques are used in applied statistics which are applicable to specific problems of real-life scenarios. Techniques used may include sample survey, quality control, index numbers etc.
4. **Inductive Statistics**: Those methods and techniques are covered here which are used to identify a specific phenomenon based on random observation. Techniques used may include Extrapolation.
5. **Analytical Statistics**: Analytical statistics uses such methods and techniques that are helpful in setting up functional relationship amidst variables. In this correlation, regression, association and attributes techniques are used.
6. **Mathematical Statistics**: It deals with the application of different mathematical theories and techniques to develop different statistical techniques. It uses techniques like integration, differentiation, trigonometry, matrix, etc.

1.7 MERITS OF STATISTICS

Following are the merits of Statistics

1. It helps to present bulk data in a precise and definite form.
2. Statistics helps to compare data and draw conclusions.
3. Statistics helps in forecasting trends.
4. Statistical data provides a base for decision making and formulating policies.

1.8 USES OF STATISTICS IN BUSINESS DECISION MAKING

The following are the main uses of statistics in various business activities:

- With the help of statistical methods, quantitative information about production, sale, purchase, finance, etc. can be obtained. This type of information helps the businessmen in formulating suitable policies.

By using the techniques of time series analysis which are based on statistical methods, the businessman can predict the effect of a large number of variables with a fair degree of accuracy.

- In business decision theory, most of the statistics techniques are used in taking a business decision which helps us in doing the business without uncertainty.
- Nowadays, a large part of modern business is being organised around systems of statistical analysis and control.. By using 'Bayesian Decision Theory', the businessmen can select the optimal decisions for the direct evaluation of the payoff for each alternative course of action.
- Uses of Mathematics for Decision Making
- The number of defects in a roll of paper, bale of cloth, sheet of a photographic film can be judged by means of Control Chart based on Normal distribution.
- In statistical quality control, we analyse the data which are based on the principles involved in Normal curve.

Uses of Statistics in Economics

Statistics is the basis of economics. The consumer's maximum satisfaction can be determined on the basis of data pertaining to income and expenditure. The various laws of demand depend on the data concerning price and quantity. The price of a commodity is well determined on the basis of data relating to its buyers, sellers, etc.

1.9 IMPORTANCE OF STATISTICS

Statistics in today's life has become an essential part of various business activities which is clear from the following points.

The importance of statistics in the following major areas:

1. Importance of Statistics in Business and Industry
2. Importance in the Field of Science and Research
3. Importance in the Field of Banking
4. Importance to the State
5. Importance in planning

Importance of Statistics in Business and Industry: In past days, decisions regarding business were made only on personal judgement. However, in these days, they are based on several mathematical and statistical techniques and the best decision is arrived by using all these techniques. For example, by using the testing hypothesis, we can reject or accept the null hypothesis which are based upon the assumption made from the population or universe. By using 'Bayesian Decision Theory' or 'Decision Theory', we can select the optimal decisions for the direct evaluation of the payoff for each alternative course of action.

Mathematics and statistics have become ingredients of various decisions problems which is clear from the following:

- **In Selecting Alternative Course of Action:** The process of business decisions involve the selection of a single action among some set of alternative actions. When there are two or more

alternative courses of action, and we need only one course of action, statistical decisions theory helps us in selecting the required course of action by applying Bayesian decision theory and thus saves lot of time.

- In Removing Uncertainty: In decision-making problems, uncertainty is very common in a situation, when the course of action is not known to us. When there are many possible outcomes of an event, we cannot predict with certainty that what will happen. By applying the concept of joint and conditional probability, the uncertainty about the event can be removed very easily.
- **In Calculating E.O.L., C.O.L., etc.:** In business, the opportunity loss is very often, which can be defined as the difference between the highest possible profit for an event and the actual profit obtained for the actual action taken. The expected opportunity loss (E.O.L.) and conditional opportunity loss (C.O.L.) can be easily calculated by using the concept of maximum and minimum criteria of pay-off.

Importance in the Field of Science and Research: Statistics has great significance in the field of physical and natural sciences. It is widely used in verifying scientific laws and phenomenon. For example, to formulate standards of body temperature, pulse rate, blood pressure, etc. The success of modern computers depends on the conclusions drawn on the basis of statistics.

Importance in the Field of Banking: In banking industry, the bankers have to relate demand deposits, time deposits, credit etc. It is on the basis of data relating to demand and time deposits that the bankers determine the credit policies. The credit policies are based on the theory of probability.

Importance to the State: We know that the subject of statistics originated for helping the ancient rulers in the assessment of their military and economic strength. Gradually its scope was enlarged to tackle other problems relating to political activities of the State.

In the modern era, the role of State has increased and various governments of the world also take care of the welfare of its people. Therefore, these governments require much greater information in the form of numerical figures for the fulfilment of welfare objectives in addition to the efficient running of their administration.

Importance in planning: Planning is indispensable for achieving faster rate of growth through the best use of a nation's resources. It also requires a good deal of statistical data on various aspects of the economy.

One of the aims of planning could be to achieve a specified rate of growth of the economy. Using statistical techniques, it is possible to assess the amounts of various resources available in the economy and accordingly determine whether the specified rate of growth is sustainable or not.

1.10 SCOPE OF STATISTICS

The following are the main scope of statistics:

1. **Presents facts in numerical figures:** The first function of statistics is to present a given problem in terms of numerical figures. We know that the numerical presentation helps in having a better understanding of nature of problem.
2. **Presents complex facts in a simplified form:** Generally, a problem to be investigated is represented by a large mass of numerical figures which are very difficult to understand and remember. Using various statistical methods, this large mass of data can be presented in a

simplified form.

3. **Studies relationship between two or more phenomena:** Statistics can be used to investigate whether two or more phenomena are related. For example, the relationship between income and consumption, demand and supply, etc.

4. **Helps in the formulation of policies:** Statistical analysis of data is the starting point in the formulation of policies in various economic, business and government activities. For example, using statistical techniques a firm can know the tastes and preferences of the consumers and decide to make its product accordingly.

5. **Helps in forecasting:** The success of planning by the Government or of a business depends to a large extent upon the accuracy of their forecasts. Statistics provides a scientific basis for making such forecasts.

6. **Provides techniques for testing of hypothesis:** A hypothesis is a statement about some characteristics of a population (or universe).

7. **Provides techniques for making decisions under uncertainty**: Many times we face an uncertain situation where any one of the many alternatives may be adopted. A businessman might face a situation of uncertain investment opportunities in which he can lose or gain. He may be interested in knowing whether to undertake a particular investment or not. The answer to such problems are provided by the statistical techniques of decision-making under uncertainty.

1.11 LIMITATIONS OF STATISTICS

Statistics is considered to be a science as well as an art, which is used as an instrument of research in almost every sphere of our activities.

Some of the limitations of statistics are as follows:

1. **Statistics Suits to the Study of Quantitative Data Only**: Statistics deals with the study of quantitative data only. By using the methods of statistics, the problems regarding production, income, price, wage, height, weight etc. can be studied. Such characteristics are quantitative in nature. The characteristics like honesty, goodwill, duty, character, beauty, intelligence, efficiency, integrity etc. are not capable of quantitative measurement and hence cannot be directly dealt with statistical methods. These characteristics are qualitative in nature. In such type of characteristics, only comparison is possible The use of statistical methods is limited to quantitative characteristics and those qualitative characteristics which are capable of being expressed numerically.

2. **Statistical Results are not Exact** :The task of statistical analysis is performed under certain conditions. It is not always possible, rather not advisable, to consider the entire population during statistical investigations. The use of samples is called for in statistical investigations. And the results obtained by using samples may not be universally true for the entire population. Data collected for a statistical enquiry may not be hundred percent true. Statistical results are true on an average.

3. **Statistics Deals with Aggregates Only**: Statistics does not recognise individual items. Consider the statement, “The weight of Mr X in the college is 70 kg”. This statement does not constitute statistical data. Statistical methods are not going to investigate anything about this statement. Whereas, if the weights of all the students of the college are given, the statistical methods may be applied to analyse that data.

4. **Statistics is Useful for Experts Only**: Statistics is both a science and an art. It is systematic and finds applications in studying problems in Economics, Business, Astronomy, Physics, Medicines etc. Statistical methods are sophisticated in nature. Everyone is not expected to possess the intelligence required to understand and to apply these methods to practical problems. This is the job of an expert, who is well-versed with statistical methods

5. **Statistics does not Provide Solutions to the Problems:** The statistical methods are used to explore the essentials of problems. It does not find use in inventing solutions to problems. For example, the methods of statistics may reveal the fact that the average result of a particular class in a college is deteriorating for the last ten years, i.e., the trend of the result is downward, but statistics cannot provide solution to this problem.

1.12 SUMMARY

- Statistics is a set of decision-making techniques which helps businessmen in making suitable policies from the available data.
- "Statistics may be defined as the collection, presentation, analysis and interpretation of numerical data."
- With the help of statistical methods, quantitative information about production, sale, purchase, finance etc. can be obtained.
- Statistics is the basis of economics. The consumer's maximum satisfaction can be determined on the basis of data pertaining to income and expenditure.
- By using 'Bayesian Decision Theory' or 'Decision Theory', we can select the optimal decisions for the direct evaluation of the payoff for each alternative course of action.
- Statistics is considered to be a science as well as an art, which is used as an instrument of research in almost every sphere of our activities.
- W.I. King says, "Statistics is a most useful servant but only of great value to those who understand its proper use."
- Yule and Kendall has rightly said that "statistical methods are most dangerous tools in the hands of inexperts."
- The statistical methods are used to explore the essentials of problems.

1.13 TEST YOUR KNOWLEDGE

1. Define statistics what are its uses of Statistics in Business Decision Making?
2. Highlight the scope of Statistics?
3. Explain the limitations of Statistics?
4. What is the scope of statistics in modern business?
5. Explain the characteristics of statistics?
6. What are the functions of statistics?
7. With the help of a chart show the limitation, function and characteristics of statistics?

8. Write an essay of uses of statistics in business decision making?

1.14 FURTHER READINGS

1. Sankalp Gaurav, Business Statistics, Agra Book International
2. Alhance D.N. and Alhance Beena, Fundamental of Statistics.
3. Monga, G.S., Elementary Statistics.
4. Gupta, S.B., Principles of Statistics.

UNIT 2 MEASURES OF CENTRAL TENDENCY MEASURES

UNIT STRUCTURE

2.0 OBJECTIVES: After reading this unit you know about

- Averages, arithmetic mean, and weighted arithmetic mean.
- Have a primary idea of median and mode
- How to calculate geometric mean and harmonic mean

2.1 INTRODUCTION

Human mind is incapable of remembering the entire mass of unwieldy data. Having learnt the methods of collection and presentation of data, one has to data. The characteristics of the data set is explored with some numerical measures namely measures of central tendency, measures of dispersion, measures of skewness, and measures of kurtosis. Here we will understand the concept of "Measure of central tendency". The measures of central tendency are also called "the averages".

In practical situations one need to have a single value to represent each variable in the whole set of data. Because, the values of the variable are not equal, however there is a general tendency of such observations to cluster around a particular level. In this situation it may be preferable to characterize each group of observations by a single value such that all other values clustered around it. That is why such measure is called the measure of central tendency of that group. A measure of central tendency is a representative value of the entire group of data. It describes the characteristic of the entire mass of data. It reduces the complexity of data and makes them amenable for the application of mathematical techniques involved in analysis and interpretation of data.

2.2 DEFINITION

The word average or the term measures of central tendency have been defined by various authors in their own way. Some of the definitions are given below :—

Simpson and Kafka observe that "A measure of central tendency is a typical value around which other figures congregate."

Lawrence J. Kaplan has defined these terms in the following words :

"One of the most widely used set of summary figures is known as measures of location, which are often referred to as averages, central tendency or central location. The purpose for computing an average value for a set of observations is to obtain a single value which is representative of all the items and which the mind can grasp in simple manner quickly. The single value is the point of location around which the individual items cluster."

Va-Lun Chou states that "an average is a typical value in the sense that it is, sometimes, employed to represent all the individual values in a series or of a variable."

It is, thus, clear from the above definitions that an average is a single value which represents a whole series and is supposed to contain its major characteristics.

2.3 OBJECTIVES OF AVERAGE

1. To get a single value which is representative of the characteristics of the entire mass of data : Averages give a bird's-eye view of the huge mass of statistical data which ordinarily are not easily intelligible.

They are devices to aid the human mind in grasping the true significance of large aggregates of facts and measurements. They set aside the unnecessary details of the data and put forward a concise picture of the complex phenomena under investigation because the human mind is not capable of grasping all the details of large numbers and their interrelationship. For example, it is not possible to keep in mind, the details of heights. weights. incomes and expenditures of even 200 students, what to talk of big figures. This difficulty of keeping all the details in mind necessitates the use of averages. An average is a single number representing the whole data and is useful in grasping the central theme of the data.

Why is an average a representative? The reason why an average is a valid representative of a series lies in the fact that ordinarily most of the items of a series cluster in the middle. On the extreme ends, the number of items is very little. In a population of 10,000 adults, there would hardly be any person who is 60 cms. high or whose height is above 240 cms. There will be a small range within which these values would vary, say, 150 cms. to 200 cms. Even within this range, a large number of persons would have a height between, say, 160 cms. and 180 cms. In other class intervals of height, the number of persons would be comparatively small. Under such circumstances if we conclude that the height of this particular group of persons would be represented by say 170 cms., we can reasonably be sure that this figure would, for all practical purposes, give us a satisfactory conclusion. This average would satisfactorily represent the whole group of figures from which it has been calculated.

2. To facilitate comparison : Since measures of central tendency or averages reduce the mass of statistical data to a single figure, they are very helpful for purposes of making comparative studies, for example, the average marks obtained by two sections of a class would give a reasonably clear picture about the level of their performance, which would not be possible if we had two full series of

marks of individual students of the two sections.

However, when such a comparison is made, we have to be careful in drawing inferences, as the marks of students in one section may vary within a small range and in the other section some students may have got very high marks and others very few marks. The comparison of averages in such a case may give misleading conclusions .

2.4 CHARACTERISTICS OF GOOD AVERAGE

1. It should be rigidly defined: If an average is left to the estimation of an observer and if it is not a definite and fixed value, it cannot be representative of a series. The bias of the investigator in such cases would considerably affect the value of the average. If the average is rigidly defined, this instability in its value would be no more, and it would always be a definite figure.

2. It should be based on all the observations of the series: If some of the items of the series are not taken into account in its calculation, the average cannot be said to be a representative one.

3. It should be capable of further algebraic treatment: If an average dose not possess this quality, its use is bound to be very limited. It will not be possible to calculate, say, the combined average of two or more series from their individual averages; further, it will not be possible to study the average relationship of various parts of a variable if it is expressed as the sum of two or more variables etc.

4. If should be easy to calculate and simple to follow: If the calculation of the average involves tedious mathematical processes, it will not be readily understood by a person of ordinary intelligence and its use will be confined only to a limited number of persons and, hence, can never be a popular measure. As such, one of the qualities of a good average is that it should not be too abstract or mathematical and should be easy to calculate.

5. It should not be affected by fluctuations of sampling: If two independent sample studies are made in any particular field, the averages thus obtained, should not materially differ from each other. No doubt, when two separate enquiries are made, there is bound to be a difference in the average values calculated but, in some cases, this difference would be great while in others comparatively less. These averages in which this difference. which is technically called "fluctuation of sampling", is less, are considered better than those in which its difference is more.

2.5 TYPES OF AVERAGES

Measures of central tendency or averages are usually of the following types :

1. **Mathematical Averages:**
   - (a) Arithmetic Average or Mean
   - (b) Geometric Mean
   - (c) Harmonic Mean
2. **Averages of Position (Positional averages) :**
   - (a) Median
   - (b) Mode

Of the above mentioned five important averages, Arithmetic Average, Median and Mode are the most popular ones. Geometric mean and Harmonic mean come next. We shall study them in this very order.

2.6 ARITHMETIC MEAN

Arithmetic Average or Mean of a series is the figure obtained by dividing the sum of the values of the various items by their number. If the heights of a group of eleven persons are 164, 169, 163, 160, 165, 168, 162, 167, 170, 166, 161 centimetres, then to find the arithmetic average of the heights of these persons we shall add these figures and divide the total so obtained, by the number of items which is 11. The total of the items in this case is 1815[1] cms. and if it is divided by 11, we get the figure of 165 cms. This is the mean or arithmetic average of the series.

Calculation of the arithmetic average in a series of individual observations

A.M. = $\frac{\text{Sum of the values}}{\text{Number of the values}} \Rightarrow$ (AM) Number of values = Sum of the values

Suppose the values of a variable are respectively X_1, X_2, X_3, ... X_n, and their arithmetic average is represented by $\overline{X}$, then

$$\overline{X} = \frac{1}{N}(X_1 + X_2 + X_3 \ldots X_n)$$

or $$\overline{X} = \frac{1}{N}\sum_{i=1}^{N} X_1 \quad \text{or} \quad \overline{X} = \frac{\sum_{i=1}^{N} X_1}{N}$$

where, $\overline{X}$ = Arithmetic average; X_1's = values of the variable; $\sum$ = Summation or total; N = Number of items.

The following example would illustrate this formula.

Example 1. Calculate the simple arithmetic average of the following items :

Size of items		
20	50	72
28	53	74
34	54	75
39	59	78
42	64	79

Solution. Direct Method :

Computation of arithmetic average Size of items

X : 20, 28, 34, 39, 42, 50, 53, 54, 59, 64, 72, 74, 75, 78, 79

$\sum X$ = 20+28+34+39+42+50+53+54+59+64+72+74+75+78+79 = 821

[1] 164 + 169 + 165 + 160 + 165 + 168 + 162 + 167 + 170 + 166 + 161 = 1815

Arithmetic mean = $\frac{\sum X}{N} = \frac{821}{15} = 54.73$

Calculation of arithmetic average in a discrete series

Direct Method: In a discrete series, the values of the variable are multiplied by their respective frequencies and the products so obtained are totalled. This total is divided by the number of items which, in a discrete series, is equal to the total of the frequencies. The resulting quotient is a simple arithmetic average of the series.

Algebraically,

If, f_1, f_2, f_3, etc., stand respectively for the frequencies of the values X_1, X_2, X_3, etc.

$$\overline{X} = \frac{1}{N}\left(X_1 f_1 + X_2 f_2 + X_3 f_3 + ... + X_n f_n\right)$$

or $\overline{X} = \frac{\sum fX}{N}$ or $\frac{\sum fX}{\sum f}$

Short-cut method I : A short-cut method can be used in the discrete series also. In this method, the deviations of the items from an assumed mean are first found out and they are multiplied by their respective frequencies. The total of these products is divided by the total frequencies and added to the assumed mean. The resulting figure is the actual arithmetic average.

Algebraically : $\overline{X} = A + \frac{\sum fdx}{N}$

where, $\sum fdx$ = the total of the products of the deviations from the assumed average and the respective frequencies of the items.

Step deviation method

In step deviation method, we define $d'x = \frac{dx}{C}$, where C is some common factor in dx values and then apply the formula :

$$\overline{X} = A + \frac{\sum f\left(\frac{dx}{C}\right)}{N} \times C.$$ This is called **short-cut method II.**

Example 2. The following table gives the marks obtained by a set of students in a certain examination. Calculate the average mark per student.

Marks	**Number of students**	**Marks**	**Number of students**
10—20	1	60—70	12
20—30	2	70—80	16

30—40	3	80—90	10
40—50	5	90—100	4
50—60	7		

Solution. Short-cut Method

Computation of average marks per student

Marks (X)	**Mid values (m.v.)**	**No. of students (f)**	**Deviation from assumed mean (55)**	**Step deviations (10) (dx)**	**Total deviation (fdx)**
10–20	15	1	– 40	– 4	– 4
20—30	25	2	– 30	– 3	– 6
30—40	35	3	– 20	– 2	– 6
40—50	45	5	– 10	– 1	– 5
50—60	55	7	– 0	0	0
60—70	65	12	+ 10	+ 1	+ 12
70—80	75	16	+ 20	+ 2	+ 32
80—90	85	10	+ 30	+ 3	+ 30
90—100	95	4	+ 40	+ 4	+ 16
		N = 60			$\sum$fdx = + 69

Arithmetic average or $\overline{X} = A + \left(\frac{\sum fdx}{N} \times i\right) = 55 + \left(\frac{69}{60} \times 10\right)$ = 66.5 marks.

Merits of arithmetic average

The arithmetic average is the most popularly used measure of central tendency. There are many reasons for its popularity. In the beginning of this chapter, we had laid down certain characteristics which an ideal average should possess. We shall now see how far the arithmetic average fulfils these conditions:

1. The first condition that an average should be rigidly defined is fulfilled by the arithmetic average. It is rigidly defined and a biased investigator shall get the same arithmetic average from the series as an unbiased one. Its value is always definite.

2. The second characteristic that an average should be based on all the observations of a series is also found in this average. Arithmetic average cannot be calculated even if a single item of a series is left out.

3. Arithmetic average is also capable of further algebraic treatment. While discussing the algebraic properties of the arithmetic average, we have already seen in detail, how various mathematical processes can be applied to it for purposes of further analysis and interpretation of data. It is on account of this characteristic of the arithmetic average that:

   (a) It is possible to find the aggregate of items of a series if only its arithmetic average and the number of items is known.

   (b) It is possible to find the arithmetic average if only the aggregate of items and their number is known.

4. The fourth characteristic laid down for an ideal average that it should be easy to calculate and simple to follow, is also found in arithmetic average. The calculation of the arithmetic average is simple and it is very easily understandable. It does not require the arraying of data which is necessary in case of some other averages. In fact, this average is so well known that to a common mean average means an arithmetic average,

   Thus, the arithmetic average

   (a) is simple to calculate,

   (b) does not need arraying of data,

   (c) is easy to understand.

5. The last characteristic of an ideal average that it should be least affected by fluctuations of sampling is also present in arithmetic average to a certain extent. If the number of items in a series is large, the arithmetic average provides a good basis of comparison, as in such cases, the abnormalities in one direction are set off against the abnormalities in the other direction.

2.7 WEIGHTED ARITHMETIC AVERAGE

Need for weighting an average : In the calculation of simple average, each item of the series is considered equally important but there may be cases where all items may not have equal importance, and some of them may be comparatively more important than the others. The fundamental purpose of finding out an average is that it shall "fairly" represent, so far as a single figure can, the central tendency of the many varying figures from which it has been calculated. This being so, it is necessary that if some items of a series are more important than others, this fact should not be overlooked altogether in the calculation of an average. If we have to find out the average income of the employees of a certain mill and if we simply add the figures of the income of the manager, an accountant, a clerk, a labourer and a watchman and divide the total by five, the average so obtained cannot be a fair representative of the income of these people. The reason is that in a mill, there may be one manager, two accountants, six clerks, one thousand labourers and one dozen watchmen, and if it is so, the relative importance of the figures of their income is not the same. Similarly, if we are finding out the change in the cost of living of a certain group of people and if we merely find the simple arithmetic average of the prices of the commodities consumed by them, the average would be unrepresentative. All the items of consumption are not equally important. The price of salt may increase by 500 per cent but this will not affect the cost of living to the extent to which it would be affected, if the price of wheat goes up only by 50%. In such cases, if an average has to maintain its representative character, it should take into account the relative importance of the different items from which it is being calculated. The simple average gives equal importance to all the items of a series.

Direct Method : In calculating the weighted arithmetic average, each value of the variable is multiplied by its weights and the products so obtained are aggregated. This total is divided by the total of weights and the resulting figure is the weighted arithmetic average.

Symbolically,

$$\overline{X}_w = \frac{X_1 w_1 + X_2 w_2 + X_3 w_3 + ... + X_n w_n}{w_1 + w_2 + w_3 + ... + w_n}$$

where $\overline{X}_w$ stands for the weighted arithmetic average, X_1, X_2, etc., for the values of the variable and w_1, w_2 etc., for their respective weights :

The formula can be written in short as :

$$\overline{X}_w = \frac{\sum Xw}{\sum w}$$

where, $\sum Xw$ stands for the sum of the products of the values and their respective weights, and $\sum w$ for the sum of the weights.

Indirect method : Weighted arithmetic mean can be calculated by an indirect method also where we assume an average and take the deviation from the assumed mean. These deviations are multiplied by respective weights of items. The sum of these products is then divided by the total of weights and added to the assumed average. This figure would be the value of the weighted arithmetic average.

Symbolically,

$$\overline{X}_w = A_w + \frac{\sum (dx) w}{\sum w}$$

where, $\overline{X}_w$ = weighted arithmetic mean; A_w = assumed mean weighted; dx = deviations of items from assumed mean; w = Weights of various items.

Example 2. Calculate simple and weighted arithmetic averages from the following data and comment

Designation	**Monthly salary (in Rs.)**	**Strength of the cadre**
Class I Officers	1,500	10
Class II Officers	800	20
Subordinate Staff	500	70
Clerical Staff	250	100
Lower Staff	100	150

Solution :

Computation of Simple and Weighted A.M.

Designation	**Monthly salary in Rs. (X)**	**Strength of the cadre (w)**	**(Xw)**
Class I Officers	1,500	10	15,000
Class II Officers	800	20	16,000
Subordinate Staff	500	70	35,000
Clerical Staff	250	100	25,000
Lower staff	100	150	15,000
N = 5	**∑X = 3,150**	**∑(w) = 350**	**∑(Xw) = 1,06,000**

Simple arithmetic average = $\frac{\sum X}{N} = \frac{3150}{5}$ = Rs. 630

Weighted arithmetic average = $\frac{\sum (Xw)}{\sum w} = \frac{1{,}06{,}000}{350}$ = Rs. 302.857

2.8 MEDIAN

Median is the value of the middle item of a series arranged in ascending or descending order of magnitude. Thus, if there are 9 items in a series arranged in ascending or descending order of magnitude, median will be the value of the 5^{th} item. This item would divide the series in two equal parts—one part containing values less than the median value and the other part containing values more than the median value. If, however, there are even number of items in a series, then there is no central item dividing the series in two equal parts. For example if there are 10 items in a series, the median value would be between the values of 5th and 6th items. It would, thus, be the arithmetic average of the values of 5th and 6th items or it would be equal to the value of the 5th item plus the value of the 6th item divided by two.

According to **A.L. Bowley,** "If the numbers of the group are ranked in order according to the measurement under consideration, then the measurement of the number most nearly one half is the median."

The above definitions of the median do not hold good in situations where a median value is surrounded by neighbouring values which are equal in magnitude to it. For example, in a series of values such as 12, 13, 14, 15, 16, 17 and 18, there is no value which is so located that three values are smaller than it and three are greater than it. However, value 15 is designated as median. Keeping in view such situations, **Croxton** and **Cowdon** have given a revised definition of median as, "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."

2.8.1 Calculation of Median

The calculation of median involves two basic steps, viz. (i) the location of the middle item and (ii) finding out its value.

The middle item in series of individual observations and also in a discrete series is $\left(\frac{n+1}{2}\right)^{th}$ item, where n is the total number of observations. In case of a continuous series $\left(\frac{n}{2}\right)^{th}$ item is the middle item of the series.

Once the middle item is located, its values has to be found out. In a series of individual observations, if the total number of items is an odd figure, the value of the middle item is the median value. If the number of items is even, the median value is the average of the two items in the centre of the distribution. The examples given below would clarify these points.

2.8.2 Computation of Median in a series of individual observations

Example 3. Find out the median of the following items :

5, 7, 9, 12, 10, 8, 7, 15, 21

Solution. These items would first be arranged in ascending order of magnitude. The series then would he as follows:

Calculation of Median

Serial Number	Size of items
1	5
2	7
3	7
4	8
5	9
6	10
7	12
8	15
9	21

If M represents the median and N the number of items.

$$M = \text{Size of } \left(\frac{n+1}{2}\right)^{th} \text{ item} = \text{Size of } \left(\frac{9+1}{2}\right)^{th} \text{ or } 5^{th} \text{ item} = 9.$$

In the above example, the number of items was odd and there was no difficulty in locating the middle items and its value. If the number of items is even, the middle item and its value would be calculated as illustrated in the following example.

Example 4. Find out the value of median from the following data :

Daily wages (R)	10	5	7	11	8
Number of Workers	15	20	15	18	12

Solution

Calculation of median

Wages in ascend-ing order (Rs.)	**Number of persons (f)**	**Cumulative Frequency (c.f.)**
5	20	20
7	15	35
8	12	47
10	15	62
11	18	80

Median is the value of $\left(\frac{N+1}{2}\right)^{th}$ or $\left(\frac{80+1}{2}\right)^{th}$ 40.5^{th} items. All items from 35 onwards upto 47 have a value of 8. Thus, the median value would be Rs. 8.

Continuous series descending order : In such series, there is a slight change in the formula for calculating median. Since the series are cumulated in descending order, the cumulative frequency of the class preceding the median class is found out by subtracting the cumulative frequency of the median class from the total of the cumulative frequency. In other words :

$$c = (N - c.f.)$$

where c = cumulative frequency less of the class preceding the median class

N = total cumulative frequency

cf = cumulative frequency of the median class.

The following example would illustrate the point :

Example 5. Calculate median from the following data :

Age	**Number of persons**	**Age**	**Number of persons**
55—60	7	35—40	30
50—55	13	30—35	33
45—50	15	25—30	28
40—45	20	20—25	14
			Total 160

Solution

Computation of Median

Age	Number of persons	Cumulative frequency
55—60	7	7
50—55	13	20
45—50	15	35
40—45	20	55
35—40	30	85
30—35	33	118
25—30	28	146
20—25	14	160

In the above example, median is the value of $\left(\frac{N}{2}\right)^{th}$ or $\left(\frac{160}{2}\right)^{th}$ or 80^{th} item which lies in 35—40 class interval.

The frequency of the class preceding the median class or the value of c = Total frequency minus the cumulative frequency of the median class or

160—85 = 75

$$\text{Median } M = l_1 + \frac{l_2 - l_1}{f_1}(m - c) = 35 + \frac{40-35}{30}(80 - 75) = 35 + \left(\frac{5}{30} \times 5\right) = 35.83$$

Example 6. Compute median from the following data :

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

Solution. Here, we are given the mid-values of the class-intervals of a continuous frequency distribution. The difference between two mid-values is 10, hence, 10/2 = 5 is reduced from each mid-value to find the lower limit and the same is added to find the upper limit of a class. The classes are, thus, 110–120, 120–130 ..... and so on upto 190–200.

Computation of Median

Class-intervals	Frequency	Cumulative frequency
110—120	6	6

120—130	25	31
130—140	48	79
140—150	72	151
150—160	116	267
160—170	60	327
170—180	38	365
180—190	22	387
190—200	3	390
Total	**390**	

The middle item is $\frac{390}{2}$ or 195 which lies in the 150—160 group.

$$M = l_1 + \frac{l_2 - l_1}{f_1}(m - c)$$

$$= 150 + \frac{160-150}{116}(195-151) = 150 + \frac{10}{116}(44)$$

$$= 153.8$$

2.9 MODE

Mode is the most common item of a series. Generally, it is the value which occurs largest number of times in a series. In the words of **Croxton** and **Cowden,** "The mode of a distribution' is the value at the point around which the items tend to be most heavily concentrated."

According to **A.M. Tuttle,** 'Mode is the value which has the greatest frequency density in its immediate neighbourhood.'

The above two definitions indicate that mode is a value around which there is the greatest concentration of values. It may not necessarily be the value which occurs the largest number of times in a series, as in some cases, the point of maximum concentration may be around some other value. In some cases, there may be more than one point of concentration of values and the series may be bi-modal or multi-modal. We shall discuss these cases later.

The word Mode is derived from the French word (*la mode*) which means fashion or the most popular phenomenon. Mode, thus, is the most popular item of a series around which there is the highest frequency density. When we speak of the 'average student', 'average collar size', 'average size of a shoe', we are referring to mode. When we say that, on an average, a student spends Rs. 300 per month, we imply that a very large number of students spend around Rs. 300 per month. It is the value of mode. It is the most typical or fashionable value of the series.

Example 7. Find the mode of the following data relating to the weight of 10 students :

Sl. No.	Weight in pounds	Sl. No.	Weight in pounds
1	20	6	130
2	130	7	132
3	135	8	132
4	130	9	135
5	140	10	141

Solution

Weight in pounds	No. of Students
120	1
130	3
132	2
135	2
140	1
141	1
	10

Since item 130 occurs the largest number of times, it is the modal value.

If there are more than one point of concentration, mode cannot be found and the series is called bi-modal.

Grouping method: In discrete and continuous series, if the items concentrate at more than one value, attempts are made to find out the point of maximum concentration with the help of grouping method. In this method, values are first arranged in ascending order and the frequencies against each value are written down. These frequencies are then added in two's and the totals are written in lines between the values added.

Frequencies can be added in two's in two ways :

(i) By adding frequencies of items number 1 and 2; 3 and 4; 5 and 6 and so on.

(ii) By adding frequencies of items number 2 and 3; 4 and 5; 6 and 7 and so on. After this, the frequencies are added in three's. This can be done in three ways:

(a) By adding frequencies of items number 1, 2 and 3, 4, 5 and 6, 7, 8 and 9 and so on.

(b) By adding frequencies of items number 2, 3 and 4, 5, 6 and 7, 8, 9, and 10 and so on.

(c) By adding the frequencies of items number 3, 4 and 5, 6, 7 and 8, 9, 10 and 11 and so on.

If necessary, frequencies can be added in four's and five's also. After this, the size of items containing the maximum frequencies are noted down and the item which has the maximum frequency the largest number of times is called the mode. If grouping has been done in case of continuous series we shall be in a position to determine the modal class by this process.

Example 8. Find the mode of the following series :

Size	Frequency	Size	Frequency
5	48	13	52
6	52	14	41
7	56	15	57
8	60	16	63
9	63	17	52
10	57	18	48
11	55	19	40
12	50	—	—

Solution

Location of mode by grouping

Size of Item (x)	Frequency (f)					
	(1)	(2)	(3)	(4)	(5)	(6)
5	48					
6	52	100				
7	56		108	156		
8	60	116			168	
9	63		123			179
10	57	120		180		
11	55		112		175	
12	50	105				162
13	52		102	157		
14	41	93			143	
15	57		98			
16	63	120		161		150
17	52		115		172	
18	48	100		140		
			88			163
19	40					

The frequencies in column (1) are first added in two's in columns (2) and (3). Then they are added in three's in columns (4), (5) and (6). The maximum frequency in each column is indicated by thick letters. It will be observed that mode changes with the change in grouping. Thus, according to column (I), mode should be 9 or 16. To find out the point of maximum concentration, the data can be arranged in the shape of table as follows:

Analysis Table

Columns	**Size of item containing maximum frequency**						
(1)			9				16
(2)			9	10		15	16
(3)		8	9				
(4)		8	9	10			
(5)			9	10	11		
(6)	7	8	9				
No. of times a size occurs	1	3	9	3	1	1	2

Since the size 9 occurs the largest number of times, it is the modal size or mode is 9.

Example 9. Find the mode from the following data :

Values	**Frequency**	**Values**	**Frequency**
Below 50	97	Below 30	60
Below 45	95	Below 25	30
Below 40	90	Below 20	12
Below 35	80	Below 15	4

Solution. The cumulative series would first be converted into a simple continuous series as follows :

Values	**Frequency**	**Values**	**Frequency**
45—50	2	25—30	30 f_1
40—45	5	20—25	18 f_0
35—40	10	15—20	8
30—35	20 f_2	10—15	4

This series does not need grouping as modal class is very prominent. The maximum frequency

30 is against the class-interval 25—30 which is the modal class. Grouping would also give the same result. Hence,

$$Z = l_1 + \frac{f_1 - f_0}{2f_1 - f_0 - f_2}(l_2 - l_1)$$

$$= 25 + \frac{30-18}{60-18-20}(30-25)$$

$$= 25 + \left(\frac{12}{22} \times 5\right) = 27.72$$

Example 10. Modal marks for a group of 94 students are 54. Ten students got marks between 0—20, thirty students between 40—60 and fourteen students between 80—100. Find out the number of students getting marks between 20—40 and 60—80 if the maximum mark of the test were 100.

Solution.

Marks	**No. of Students**
0—20	10
20—40	x
40—60	30
60—80	y
80—100	14
	94

Mode = $l_1 + \frac{f_1 - f_0}{2f_1 - f_0 - f_2}(l_2 - l_1)$, Mode is given as 54

so $$54 = 40 + \frac{30 - x}{60 - x - y}(60 - 40)$$

or $$14 = \frac{30 - x}{60 - x - y} \times 20 = \frac{600 - 20x}{60 - x - y}$$

or $840 - 14x - 14y = 600 - 20x$ or $6x - 14y = -240$

The total number of students is 94. Therefore, the missing values (x + y) would be (94 – 10 – 30 – 14) or 40.

So, we have two equations :

$6x - 14y = -240$ or $x + y = 40$

If they are solved as simultaneous equation, we get :

$6x - 14y = -240$ ... (i)

$6x + 6y = 240$ ... (ii)

Subtracting equation (ii) from (i) we get :

$-20y = -480$ or $y = 24$

Since $x + y = 40$, therefore $x = 40 - 24$ or 16.

The missing values are, thus, 16 and 24.

2.10 GEOMETRIC MEAN

Geometric mean is defined as the n^{th} root of the product of n items of a series. Thus, if the geometric mean of 3, 6 and 8 is to be calculated it would be equal to the cube root of the product of these figures. Similarly, the geometric mean of 8, 9, 12 and 16 would be the 4^{th} root of the product of these four figures.

Symbolically, GM = $\sqrt[n]{X_1.X_2.X_3....X_n}$

where GM stands for the geometric mean, N for the number of items and X for the values of the variable.

The calculation of the geometric mean by this process is possible only if the number of items is very few. If the number of items is large and their size is big, this method is more or less out of question. In such cases, calculations have to be done with the help of logarithm. In terms of logs.

$$GM = \left[X_1.X_2.X_3....X_n\right]^{\frac{1}{n}}$$

$$\Rightarrow \quad \log GM = \frac{\log X_1 + \log X_2 + \log X_3....\log X_n}{N}$$

$$GM = \text{Anti-log}\left[\frac{\log X_1 + \log X_2 + \log X_3....\log X_n}{N}\right]$$

or

$$GM = \text{Anti-log}\left[\frac{\Sigma \log X}{N}\right]$$

Thus, geometric mean is the anti-log of the arithmetic average of the logs of the values of a variable. It should be noted that the value of the geometric mean is always less than the value of the arithmetic average unless all the items have equal value in which case the geometric mean and arithmetic average have identical values.

The following examples would illustrate the calculation of geometric mean.

Calculation of geometric mean in a series of individual observations.

Example 11. Calculate the simple geometric mean from the following items:

133, 141, 125, 173, 182

Solution.

Calculation of the geometric mean

Size of item	Logarithms
133	2.1239
141	2.1492
125	2.0969
173	2.2380
183	2.2601
N = 5	**$\sum$ log s = 10.8681**

According to the formula, viz.,

Geometric Mean = $\sqrt[n]{X_1.X_2.X_3....X_n}$

$$GM = \text{Anti-log}\left[\frac{\log X_1 + \log X_2 + \log X_3 .... \log X_n}{N}\right]$$

$$GM = \text{Anti-log}\left(\frac{10.8681}{5}\right) = \text{Anti-log } 2{,}1736$$

GM = 149 (to the nearest whole number)

Thus, the geometric mean is 149.

Geometric Mean can also be calculated by assuming the logarithm of a number and taking deviations of' the logs of actual items from the assumed log. In such a case,

$$GM = \text{Anti-log}\left[\text{assumed log} + \frac{\sum \text{Deviations}}{N}\right]$$

The earlier example has been solved in this manner below:

Alternate Method

Size of item (X)	logs X	Deviation from assumed log mean (2.000) dx
133	2.1239	.1239
141	2.1492	.1492
125	2.0969	.0969

173	2.2380	.2380
182	2.2601	.2601
N = 5		**$\sum$ log dx = .8681**

Geometric Mean = Anti-log $\left[\text{assumed log} + \frac{\sum \text{Deviations}}{N}\right]$

= Anti-log $\left[2 + \frac{08681}{5}\right]$ = Anti-log 2.1736.

= 149 (to the nearest whole number)

Thus, the geometric mean is 149.

Calculation of GM in discrete series

In discrete series, the geometric mean of

$$GM = \text{Anti-log } \frac{\sum f \log X}{N}$$

where, f is the frequency, X the value of the item and N the total number of items.

The steps of calculation are :

(i) Find the logarithms of variable X.

(ii) Multiply these logs with the respective frequencies, and total all such values, it would be $\sum$ f log X.

(iii) Divide $\sum$ f log X by total frequency or N.

(iv) Find out the anti-log of this value $\frac{\sum f \log X}{N}$. It will be the value of the geometric mean.

Example 12. Find the geometric mean of the following distribution :

Values	**Frequency**
352	48
220	10
230	8
160	12
190	15

Solution.

Values (X)	Frequency (f)	log X	flog X
352	48	2.5465	122.2320
220	10	2.3424	23.4240
230	8	2.3617	18.8936
160	12	2.2041	26.4492
190	15	2.2788	34.1820
Total	N = 93		$\sum$ f log X = 225.1808

Geometric Mean or GM = Anti-log $\left[\frac{\sum f \log X}{N}\right]$

= Anti-log $\frac{(225.1808)}{93}$ = Anti-log 2.4134 = 263.8.

2.11 HARMONIC MEAN

Harmonic Mean of a series is the reciprocal of the arithmetic average of the reciprocal of the values of its various items.

Symbolically, the Harmonic mean or HM of a series :

$$HM = \text{Reciprocal of } \frac{\frac{1}{X_1}+\frac{1}{X_2}+....+\frac{1}{X_n}}{N}$$

$$= \frac{1}{\frac{\frac{1}{X_1}+\frac{1}{X_2}+....+\frac{1}{X_n}}{N}} = \frac{1}{\frac{1}{X_1}+\frac{1}{X_2}+....+\frac{1}{X_n}}$$

where X stands for value of the variable.

Thus, the Harmonic Mean of 2, 4 and 8 would be reciprocal of $\frac{\frac{1}{2}+\frac{1}{4}+\frac{1}{8}}{3}$ or reciprocal of $\frac{\frac{7}{8}}{3}$ or reciprocal of $\frac{7}{24}$ or $\frac{24}{7}$ = 3.43.

If the number of items in a series is large it would be a tedious job to find the reciprocal of each item and then to total them and then to divide the total by the number of items and then to find out the

reciprocal of the value. The formula can be simplified in the following manner :

$$\text{HM} = \text{Reciprocal of } \frac{\frac{1}{X_1}+\frac{1}{X_2}+\ldots+\frac{1}{X_n}}{N}$$

or $$\frac{N}{\frac{1}{X_1}+\frac{1}{X_2}+\ldots+\frac{1}{X_n}} \quad \text{or} \quad \frac{N}{\Sigma\left(\frac{1}{X}\right)}$$

Thus, H.M. is equal to the number of items or N divided by the sum of the reciprocals of different items. This formula is used in the series of individual observation.

In discrete series: We first find out the reciprocal of each value and multiply it by the concerned frequency. Then total the products and divide the total by the total frequency and then find out the reciprocal of the value, which is the harmonic mean of the series. Thus, in a discrete series:

$$\text{H.M.} = \text{Reciprocal of } \frac{f_1\left(\frac{1}{X_1}\right)+f_2\left(\frac{1}{X_2}\right)+\ldots.+f_n\left(\frac{1}{X_n}\right)}{\Sigma f}$$

$$= \frac{\Sigma f}{f_1\left(\frac{1}{X_1}\right)+f_2\left(\frac{1}{X_2}\right)+\ldots.+f_n\left(\frac{1}{X_n}\right)}$$

This formula can again be simplified as :

$$\text{H.M.} = \frac{\Sigma f}{\Sigma\left(f\times\frac{1}{X}\right)} = \frac{N}{\Sigma\left(f\times\frac{1}{X}\right)}$$

In continuous series, the value of the variable, is the mid-point of the class-interval or m.v. = X and the formula for the calculation of H.M. in such a series would be :

$$\text{H.M.} = \frac{\Sigma f}{\Sigma\left(f\times\frac{1}{M}\right)} = \frac{N}{\Sigma\left(f\times\frac{1}{M}\right)}$$

We shall now use these formula to calculate the H.M. in the foregoing examples.

Calculation of Harmonic in a series of individual observations.

Example 13. Calculate the Harmonic mean of the following values:

15, 250, 15.7, 157, 1.57, 105.7, 10.5, 1.06, 25.7 and 0.257

Solution. We shall find out the reciprocals of the above values from the mathematical tables which are available, instead of manual calculation.

Calculation of Harmonic Mean

Values (X)	Reciprocal (1/X)
15	0.06667
250	0.00400
15.7	0.06369
157	0.00637
1.57	0.63690
105.7	0.00946
10.5	0.09524
1.06	0.94340
25.7	0.03891
0.257	3.89100
N = 10	5.75564 $\sum(1/X)$

$$\text{H.M.} = \frac{N}{\sum\left(\frac{1}{X}\right)} = \frac{10}{5.75564} = 1.735$$

Calculation of H.M. in continuous series

As has been pointed out earlier, in a continuous series, the mid-values of class-interval represent the value of the variable. Once this is done, the continuous series becomes a discrete series and the HM is easily calculate.

Example 14. From the following data, calculate Harmonic Mean:

Class-interval :	10-20	20-30	30-40	40-50	50-60
Frequency :	30	75	70	135	220

Solution.

Calculation of Harmonic Mean

Class Interval (X)	Frequency (f)	Mid-value (m.v.)=m	f/m
10–20	30	15	2
20–30	75	25	3

30–40	70	35	2
40–50	135	45	3
50–60	220	55	4
	N = 530		$\sum$ (f/m) = 14

$$\text{H.M.} = \frac{N}{\sum\left(\frac{f}{m}\right)} = \frac{530}{14} = 37.86.$$

2.12 SUMMARY

Mean, median, and mode are the three measures of central tendency in statistics. We identify the central position of any data set while describing a set of data. This is known as the measure of central tendency. We come across data every day. We find them in newspapers, articles, in our bank statements, mobile and electricity bills. The list is endless; they are present all around us. Now the question arises if we can figure out some important features of the data by considering only certain representatives of the data. This is possible by using measures of central tendency or averages, namely mean, median, and mode. Mean, median, and mode are the measures of central tendency, used to study the various characteristics of a given set of data. A **measure of central tendency** describes a set of data by identifying the central position in the data set as a single value. We can think of it as a tendency of data to cluster around a middle value. In statistics, the three most common measures of central tendencies are Mean, Median, and Mode.

- **Mean:** The mean is also known as the average, and it is calculated by adding up all the values in a data set and dividing by the total number of values.
- **Median:** The median is the middle value of a data set, which separates the highest and lowest values equally. It is calculated by arranging the data set in order from lowest to highest and finding the value in the exact middle.
- **Mode:** The mode is the value that appears most frequently in a data set.

2.13 TEST YOUR KNOWLEDGE

1. What is mean? Explain its characteristics
2. What is mode?
3. What is median?
4. Explain the geometric mean?
5. Explain harmonic mean?

Numerical Questions

1. Find the mean of first 5 prime numbers.

Ans 5.6

2. Find the mean of first ten whole numbers.

Ans. 4.5

3. Find the mean of the following data.

(a) 9, 7, 11, 13, 2, 4, 5, 5

(b) 16, 18, 19, 21, 23, 23, 27, 29, 29, 35

(c) 2.2, 10.2, 14.7, 5.9, 4.9, 11.1, 10.5 (d) $1^1/_4$, $2^1/_2$, $5^1/_2$, $3^1/_4$, $2^1/_2$

Ans (a) 7, (b) 24, (c) 8.5, (d) 3

4. The mean of 8, 11, 6, 14, x and 13 is 66. Find the value of the observation x.

Ans. 344

5. The mean of 6, 8, x + 2, 10, 2x - 1, and 2 is 9. Find the value of x and also the value of the observation in the data.

Ans. 9, 11, 17

6. Find the mean of the following distribution.

x_i	1	2	3	4	5
f_i	4	5	8	10	3

Ans 3.1

2.14 Further Readings

1. Sankalp Gaurav, Business Statistics, Agra book International , Agra
2. Sinha, V.C., Principles of Statistics.
3. Gupta, S.B., Principles of Statistics.
4. Monga, G.S., Elementary Statistics.

UNIT 3 - MEASURES OF DISPERSION

UNIT STRUCTURE

1. Introduction
2. Definition
3. Objectives of Measuring Dispersion
4. Different Measures of Dispersion
5. Range
6. Standard Deviation
7. Lorenz Curve
8. Skewness
9. Kurtosis
10. Summary
11. Test your knowledge
12. Further Readings

3.0 OBJECTIVES

After going through this unit you should be able to know about the–

1. Dispersion
2. Range
3. Standard Deviation
4. skweness
5. Kurtosis

3.1 **Introduction**

Averages fail to reveal the full details of the distribution. Two or three distributions may have the same average but still they may differ from each other in many ways. In such cases, rather statistical analysis of the data is necessary so that these differences between various series can be studied and accounted for such analysis will make our results more accurate and we shall be more confident of our conclusions.

Suppose, there are three series of nine items each as follows:

Series A	**Series B**	**Series C**
40	36	1
40	37	9

40	38	20
40	39	30
40	40	40
40	41	50
40	42	60
40	43	70
40	44	80
Total 360	360	360
Mean 40	40	40

In the first series, the mean is 40 and the value of all the items is identical. The items are not at all scattered, and the mean fully discloses the characteristics of this distribution. However, in the second case, though the mean is 40 yet all the items of the series have different values. But the items are not very much scattered as the minimum value of the series is 36 and the maximum is 44 in the range. In this case also, mean is a good representative of the series because the difference between the mean and other items is not very significant. In the third series also, the mean is 40 and the values of different items are also different, but here the values are very widely scattered and the mean is 40 times of the smallest value of the series and half of the maximum value. Though the mean is the same in all the three series, yet the series differ widely from each other in their formation. Obviously, the average does not satisfactorily represent the individual items in this group and to know about the series completely, further analysis is essential. The scatter among the items in the first case is nil, in the second case it varies within a small range, while in the third case the values range between a very big span and they are widely scattered. It is evident from the above, that a study of the extent of the scatter around average should also be made to throw more light on the composition of a series. The name given to this scatter is dispersion.

3.2 **Definition**

Some important definitions of dispersion are given below:

(i) "Dispersion or spread is the degree of the scatter or variation of the variable about a central value." – *Brooks and Dick*

(ii) "Dispersion is the measure of the variations of the items." – *A.L. Bowley*

(iii)"The degree to which numerical data tend to spread about an average value is called the variation or dispersion of the data." – *Spiegel*

From the above definitions, it is clear that in a general sense the term dispersion refers to the variability in the size of items. If the variation is substantial, dispersion is said to be considerable and if the variation is very little, dispersion is insignificant.

Usually, in a precise study of dispersion the deviations of size of items from a measure of central tendency are found out and then these deviations are averaged to give a single figure representing the dispersion of the series. This figure can

be compared with similar figures representing other series. Such comparisons give a better idea about the formation of series than a mere comparison of their averages.

Averages of second order : For a precise study of dispersion, we have to average deviations of the values of the various items, from their average. We have seen earlier that arithmetic mean, median, mode, geometric mean and harmonic mean, etc., are all averages of the first order. Since in the calculation of measures of dispersion, the average values are derived by the use of the averages of the first order, the measures of dispersion are called averages of the second order.

3.3 Objectives of Measuring Dispersion

Measures of variations are calculated to serve the following purposes:

(i) To judge the reliability of measures of central tendency.

(ii) To make a comparative study of the variability of two series.

(iii) To identify the causes of variability with a view to control it.

Spur and **Bonimi** have very rightly observed that, "in matters of health, variations in body temperature, pulse beats and blood pressure are basic guides to diagnosis. Prescribed treatment is designed to control their variation. In industrial production, efficient operation requires control of quality variation, the causes of which are sought through inspection and quality control programs."

In Social Sciences where we have to study problems relating to inequality in income and wealth, measures of dispersion are of immense help.

(iv) To serve as a basis for further statistical analysis.

The properties of a good measure of dispersion are the same as the properties of a good measure of central tendency. Precisely, they are:

(i) It should be rigidly defined.

(ii) It should be based on all the observations of the series.

(iii) It should be capable of further algebraic treatment.

(iv) It should be easy to calculate and simple to follow.

(v) It should not be affected by fluctuations of sampling.

3.4 Different Measures of Dispersion

Absolute and relative "dispersion: Dispersion or variation can be expressed either in terms of the original units of a series or as an abstract figure like a ratio or percentage. If we calculate dispersion of a series relating to the income of a group of persons in absolute figures, it will have to be expressed in the unit in which the original data are, say, rupees. Thus, we can say that the income of a group of persons is Rs. 5000 per month and the dispersion is Rs. 500. This is called Absolute Dispersion. If, on the other hand, dispersion is measured as a percentage or ratio of the average, it is called Relative Dispersion. Since the relative dispersion is a ratio, it has no units. In the above case, the average income would be referred to as Rs. 5000 per month and the relative dispersion $\frac{500}{5000} = 0.1$ or 10%. In a comparison of the variability of two or more series, it is the relative

dispersion that has to be taken into account as the absolute dispersion may be erroneous or unfit for comparison if the series are originally in different units.

3.5 Range

Range is the simplest possible measure of dispersion. It is the difference between the values of the extreme items of a series. Thus, if in a series relating to the weight measurements of a group of students, the lightest student has a weight of 40 kg. and the heaviest, of 110 kg. The value of range would be 110 — 40 = 70 kg. This figure indicates the variability in the weight of students.

Symbolically,

$$\text{Range (R)} = L - S$$

where, L is the largest value and S the smallest value in a series.

Range as calculated above is an absolute measure of dispersion which is unfit for purposes of comparison if the distributions are in different units. For example, the range of the weights of students cannot be compared with the range of their height measurements as the range of weights would be in kg. and that of heights in centimetres. Sometimes, for purpose of comparison, a relative measure of range is calculated. If range is divided by the sum of the extreme items, the resulting figure is called "The Coefficient of the Range" or "The Coefficient of the Scatter."

Symbolically,

The Ratio of Range or the Coefficient of the scatter (or Range)

$$= \frac{\text{Max. value} - \text{Min. value}}{\text{Max. value} + \text{Min. value}} = \frac{L - S}{L + S}$$

$$= \frac{\text{Absolute range}}{\text{Sum of the extreme values}}$$

The following illustration would illustrates the use of the above formulae

Example. The profits of a company for the last 8 years are given below. Calculate the Range and its Coefficient:

Year	1975	1976	1977	1978	1979	1980	1981	1982
Profits (in '000 Rs.)	40	30	80	100	120	90	200	230

Solution.

Here, L = 230 and S = 30

Range = L – S = 230 – 30 = 200

Coefficient of Range = $\frac{L - S}{L + S}$ or $\frac{230 - 30}{230 + 30}$ or $\frac{200}{260} = 0.77$

Merits and Demerits of Range

As has been pointed out that' a good measure of dispersion should be rigidly defined, easily calculated, readily understood, should be capable of further mathematical treatment and should not be much affected by fluctuations of sampling.

Out of these the only merit possessed by Range is that it is easily calculated and readily understood.

As against this, the Range as a measure of dispersion has the following demerits :

1. It is affected greatly by fluctuation of Sampling.
2. It is not based on all the observations of the series.
3. It cannot be used in case of open distributions.
4.

Uses of Range

With all its limitations Range is commonly used in certain fields. For example:

1. Quality Control.
2. Variation in Money Sales, Share values, Exchange Rates and Gold Prices, etc.
3. Weather forecasting.

3.6 Standard Deviation

The concept of standard deviation was first used by Karl Pearson in the year 1893. It is the most commonly used measure of dispersion. It satisfies most of the properties laid down for an ideal measure of dispersion.

Meaning: The technique of the calculation of mean deviation is mathematically illogical as in its calculation, the algebraic signs are ignored. This drawback is removed in the calculation of standard deviation, where squares of the deviations from the mean one used. Standard deviation is the square root of the arithmetic average of the squares of the deviations measured from the mean. The standard deviation is conventionally represented by the Greek letter Sigma □.

Symbolically, $$\sigma = \sqrt{\frac{\Sigma(X - \overline{X})^2}{N}}$$

Where □ stands for the standard deviation,

Calculation of Standard deviation

1. Series of individual observation: In such a series the standard deviation can be calculated in any of the following ways:

Direct method No. 1 : In this method, the following steps are involved:

(i) Find the arithmetic average of the series.

(ii) Find the deviations of each item from the arithmetic average and denote it by (d) i.e., find $(X - \overline{X})$ for each X.

(iii) Square these deviations and total them to find $\sum d^2$.

(iv) Divide $\sum d^2$ by the number of items to find $\frac{\sum d^2}{N}$. This figure is called the second moment about N the Mean.

(v) Standard Deviation = VN = $\sqrt{\frac{\sum d^2}{N}} = \sqrt{\frac{\sum (X - \overline{X})^2}{N}}$

Example. Calculate the standard deviation of the heights of 10 students given below :

Heights (in cms.) : 160, 160, 161, 162, 163, 163, 163, 164, 164, 170

Solution. Calculation of Standard Deviation of heights.

Height in centimeters	**Deviations from mean 163**	**Deviations squared (d^2)**
160	– 3	9
160	– 3	9
161	– 2	4
162	– 1	1
163	0	0
163	0	0
163	0	0
164	+ 1	1
164	+ 2	1
170	+ 7	49
		$\sum d^2 = 74$

Arithmetic average or X = $\frac{\sum X}{N} = \frac{1630}{10} = 163$ Cms.

Standard Deviation or □ = $\sqrt{\frac{\sum d^2}{N}} = \sqrt{\frac{74}{10}} = \sqrt{7.4} = 2.27$ Cms.

Second method for Calculating S.D.

Calculation of Standard deviation

Discrete series. In discrete series also standard deviation can be calculated by

(i) The direct method as well as by

(ii) The short-cut method. Further it is possible to have step deviations, both in the direct and short-cut methods :

(a) Direct method:

1. Calculate the arithmetic mean.
2. Find out the deviations (d) of the various values, from the mean value. Square these deviations (d^2).
3. Multiply d^2 with the respective frequencies (f) against various values and add all such values ($\sum fd^2$).
4. Divide $\sum fd^2$ by the number of items (N) and find out the square root of the figure so obtained i.e., find $\sqrt{\frac{\sum fd^2}{N}}$. This will be the value of the Standard Deviation.

(b) Short-cut Method

1. Assume a mean (A) and take deviations (dx) from it and square them up (dx^2).
2. Multiply dx^2 with the respective frequencies or f to get (fdx^2). Total them to get ($\sum fdx^2$).
3. Divide ($\sum fdx^2$) by the number of items or N to get $\frac{\left(\sum fdx^2\right)}{N}$.
4. From $\left(\frac{\sum fdx^2}{N}\right)$ subtract the square of the difference between actual and assumed average $\left(\overline{X} - A\right)^2$ to get $\left(\frac{\sum fdx^2}{N}\right) - \left(\overline{X} - A\right)^2$.
5. Find out the square root of the above value or $\sqrt{\frac{\left(\sum fdx^2\right)}{N} - \left(\overline{X} - A\right)^2}$ and it will be the value of the standard deviation. This formula can also be written as :

$$\sigma = \sqrt{\frac{\left(\sum fdx^2\right)}{N} - \left(\frac{\sum fdx}{N}\right)^2} \quad \text{as} \quad \overline{X} - A = \frac{\sum fdx}{N}$$

Example. Calculate standard deviation for the following distribution :

Values	10	20	30	40	50	60	70
Frequency	1	5	12	22	17	9	4

Solution

Calculation of standard deviation (Step Deviation Method)

Values (X)	Freq-uency (f)	Dev. from assumed av. 40 $\frac{X-40}{10}=dx$	Total Step Dev. fdx	dx^2	fdx^2
10	1	– 3	3	9	9
20	5	– 2	– 10	4	20
30	12	– 1	– 12	1	12
40	22	0	0	0	0
50	17	+ 1	+ 17	1	17
60	9	+ 2	+ 18	4	36
70	4	+ 3	+ 12	9	36
	N = 70		$\sum fdx$ = +22		$\sum fdx^2$ =130

$$\sigma=\sqrt{\frac{\sum fdx^2}{N}-\left(\frac{\sum fdx}{N}\right)^2}\times i=\sqrt{\frac{130}{70}-\left(\frac{+22}{70}\right)^2}\times 10=\sqrt{1.757}\times 10$$

$= 1.326 \times 10 = 13.26$

Thus, the standard deviation is 13.26.

3.7 Lorenz Curve

Dispersion can be studied graphically also with the help of what is called Lorenz Curve, after the name of Dr. Lorenz who first studied the dispersion of distribution of wealth by the graphic method. The technique of drawing Lorenz Curve is not very difficult. The technique is as follows:

(i) The size of items as well as frequencies are first cumulated.

(ii) Then taking the final cumulated figure as 100 percentages are calculated for all cumulative values.

(iii) On the X-axis begin from 100 to 0 and let it represent frequencies (This is not a hard and fast rule. X-axis can begin with 0 to 100 and can represent values instead of frequencies. However, the suggested procedure gives a more convenient curve).

(iv) On the y-axis begin from 0 to 100 and let it represent values. (It can also begin from 100 to

0 and represent frequencies).

(v) Draw a line from 0 of the X-axis to the 100 of y-axis. This is the line of equal distribution.

(vi) Plot the various points of x and y and draw the curve. If the distribution is not proportionately equal, the curve would be away from the line of equal distribution. The farther is the curve from the line of equal distribution, the greater is the variability in the series.

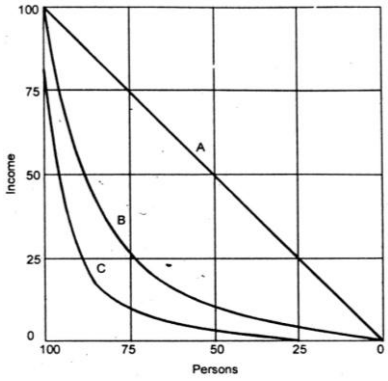
Example. Draw a Lorenz curve from the following data:

Income in thousand rupees	**Number of persons in thousands**		
	Group A	**Group B**	**Group C**
10 20 40 50 80	5 10 20 25 40	8 7 5 3 5	15 6 2 1 1

To draw the Lorenz curve from the above data, the size of the item and frequencies would have to be cumulated and then percentages would have to be calculated by taking the respective totals as 100. This has been done in the following table:

Rupees (000)	**Cumulated Income**	**Cum. %**	**No. of persons (000)**	**Cum-Number**	**Cum. %**	**No. of persons (000)**	**Cum-Number**	**Cum. %**	**No. of persons (000)**	**Cum-Number**	**Cum. %**
10 20 40 50 80	10 30 70 120 200	5 15 35 60 100	5 10 20 25 40	5 15 35 60 100	5 15 35 60 100	8 7 5 3 2	8 15 20 23 25	32 60 80 92 100	15 6 2 1 1	15 21 23 24 25	60 84 92 96 100

Now, the cumulative percentages would be plotted on a graph paper. Percentages relating to the number of persons would be shown on the abscissa and from left to right the scale would begin with 100 and end with 0. The income percentages would be shown on the ordinate and here the scale will begin with 0 at the bottom and go upto 100 at the top. The above percentages would give the following type of curve.



From the above figure, it is clear that in the first group of persons, the distribution of income is proportionately equal, so that 5% of the income is shared by 5% of the population, 15% of the income by 15% of the population and so on. It gives the line of equal distribution. In the second group, the distribution is uneven so that 5% of the income is shared by 32% of the people and 15% of the income by 60% of the people. In the third group, the distribution is still more unequal so that 5% of that income is shared by 60% of the people and 15% of the income by 84% of the people. The variation in group C is, thus, greater tnan the variation in group B. Curve C is, thus, at a greater distance from the line of equal distributions, than curve B.

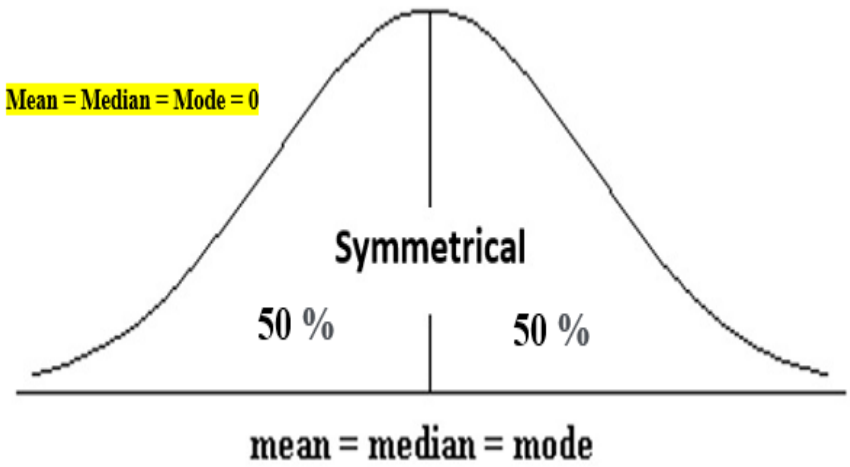
The Lorenz curve has a great drawback. It does not give any numerical value of the measure of dispersion. It merely gives a picture of the extent to which a series is pulled away from an equal distribution. It should be used along with some numerical measure of dispersion. It is very useful in the study of income distributions, distributions of land and wages, etc.

3.8 Skewness

If the values of a specific independent variable (feature) are skewed, depending on the model, skewness may violate model assumptions or may reduce the interpretation of feature importance.

In statistics, skewness is a degree of asymmetry observed in a probability distribution that deviates from the symmetrical normal distribution (bell curve) in a given set of data.

The normal distribution helps to know a skewness. When we talk about normal distribution, data symmetrically distributed. The symmetrical distribution has zero skewness as all measures of a central tendency lies in the middle.

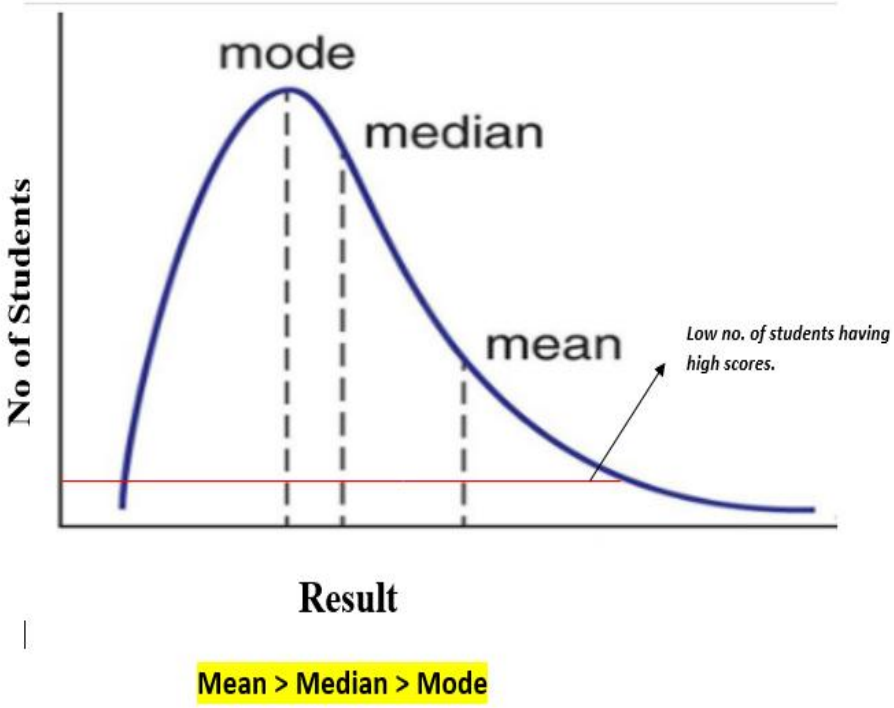


When data is symmetrically distributed, the left-hand side, and right-hand side, contain the same number of observations. (If the dataset has 90 values, then the left-hand side has 45 observations, and the right-hand side has 45 observations.). But, what if not symmetrical distributed? That data is called asymmetrical data, and that time skewness comes into the picture.

Types of skewness

1. Positive skewed or right-skewed

In statistics, a positively skewed distribution is a sort of distribution where, u*nlike symmetrically distributed data where all measures of the central tendency (mean, median, and mode) equal each other*, with positively skewed data, the measures are dispersing, which means Positively Skewed Distribution is a type of distribution where the mean, median, and mode of the distribution are positive rather than negative or zero.



Mean > Median > Mode

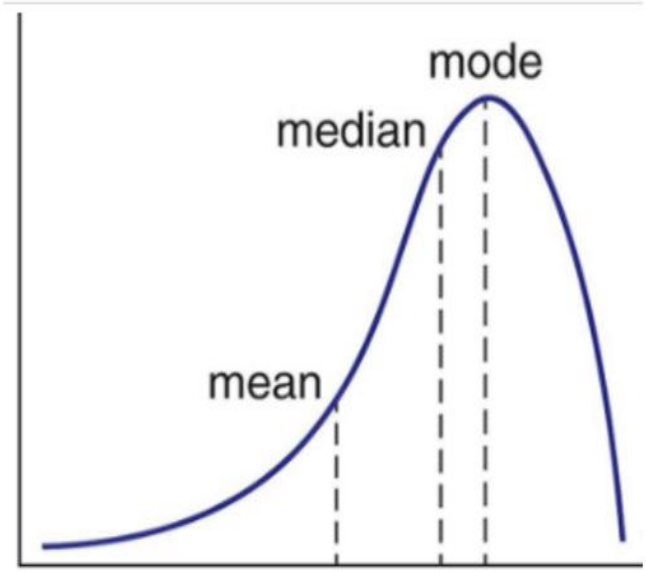
In positively skewed, the mean of the data is greater than the median (a large number of data-pushed on the right-hand side). In other words, the results are bent towards the lower side. The mean will be more than the median as the median is the middle value and mode is always the highest value

The extreme positive skewness is not desirable for distribution, as a high level of skewness can cause misleading results. The data transformation tools are helping to make the skewed data closer to a normal distribution. For positively skewed distributions, the famous transformation is the log transformation. The log transformation proposes the calculations of the natural logarithm for each value in the dataset.

2. Negative skewed or left-skewed

A negatively skewed distribution is the straight reverse of a positively skewed distribution. In statistics, negatively skewed distribution refers to the distribution model where more values are plots on the right side of the graph, and the tail of the distribution is spreading on the left side.

In negatively skewed, the mean of the data is less than the median (a large number of data-pushed on the left-hand side). Negatively Skewed Distribution is a type of distribution where the mean, median, and mode of the distribution are negative rather than positive or zero.



Median is the middle value, and mode is the highest value, and due to unbalanced distribution median will be higher than the mean.

Calculate the skewness coefficient of the sample

Pearson's first coefficient of skewness

Subtract a mode from a mean, then divides the difference by standard deviation.

$$\text{Pearson's first coefficient} = \frac{\text{Mean} - \text{Mode}}{\text{Standard Deviation}}$$

As Pearson's correlation coefficient differs from -1 (perfect negative linear relationship) to +1 (perfect positive linear relationship), including a value of 0 indicating no linear relationship, When we divide the covariance values by the standard deviation, it truly scales the value down to a limited range of **-1 to +1.** That accurately the range of the correlation values.

Pearson's first coefficient of skewness is helping if the data present high mode. But, if the data have low mode or various modes, Pearson's first coefficient is not preferred, and Pearson's second coefficient may be superior, as it does not rely on the mode.

Pearson's second coefficient of skewness

Multiply the difference by 3, and divide the product by standard deviation.

$$\text{Pearson's second coefficient} = \frac{3\,(\text{Mean} - \text{Median})}{\text{Standard Deviation}}$$

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

If the skewness is between -0.5 & 0.5, the data are nearly symmetrical.

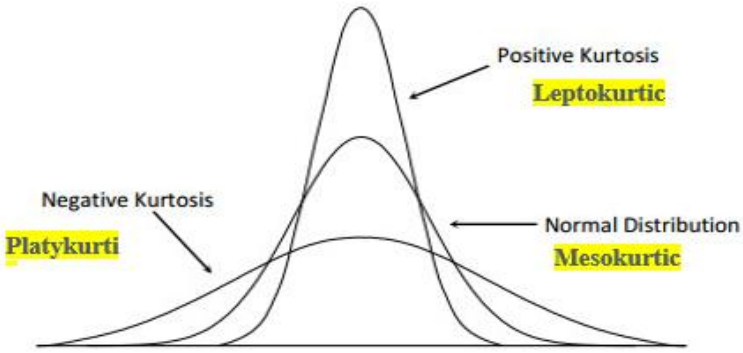
If the skewness is between -1 & -0.5 (negative skewed) or between 0.5 & 1(positive skewed), the data are slightly skewed.

If the skewness is lower than -1 (negative skewed) or greater than 1 (positive skewed), the data are extremely skewed.

3.9 Kurtosis

Kurtosis refers to the degree of presence of outliers in the distribution.

Kurtosis is a statistical measure, whether the data is heavy-tailed or light-tailed in a normal distribution.



Excess Kurtosis

The excess kurtosis is used in statistics and probability theory to compare the kurtosis coefficient with that normal distribution. Excess kurtosis can be positive (Leptokurtic distribution), negative (Platykurtic distribution), or near to zero (Mesokurtic distribution). Since normal distributions have a kurtosis of 3, excess kurtosis is calculating by subtracting kurtosis by 3.

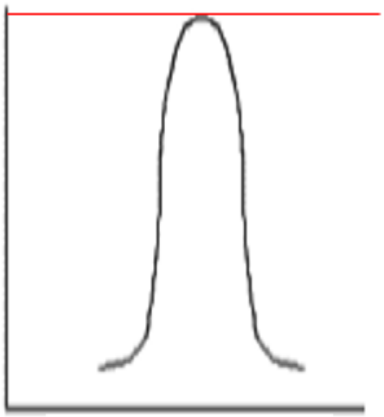
Excess kurtosis = Kurt – 3

Types of excess kurtosis

1. Leptokurtic or heavy-tailed distribution (kurtosis more than normal distribution).
2. Mesokurtic (kurtosis same as the normal distribution).
3. Platykurtic or short-tailed distribution (kurtosis less than normal distribution).

Leptokurtic (kurtosis > 3)

Leptokurtic is having very long and skinny tails, which means there are more chances of outliers. Positive values of kurtosis indicate that distribution is peaked and possesses thick tails. An extreme positive kurtosis indicates a distribution where more of the numbers are located in the tails of the distribution instead of around the mean.



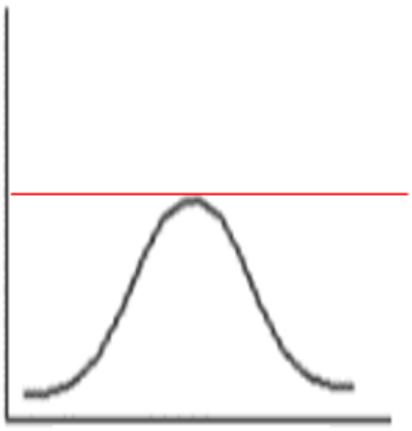
Leptokurtic

Platykurtic (kurtosis < 3)

Platykurtic having a lower tail and stretched around center tails means most of the data points are present in high proximity with mean. A platykurtic distribution is flatter (less peaked) when compared with the normal distribution.

Mesokurtic (kurtosis = 3)

Mesokurtic is the same as the normal distribution, which means kurtosis is near to 0. In Mesokurtic, distributions are moderate in breadth, and curves are a medium peaked height.



Mesokurtic

Mesokurtic = 3 – 3 = 0

3.10 Summary

Dispersion is the degree of scatter or variation of variable about a central value. In simple words Dispersion refers to the variability in the size of the items. If the values of a specific independent variable (feature) are skewed, depending on the model, skewness may violate model assumptions or may reduce the interpretation of feature importance.

In statistics, skewness is a degree of asymmetry observed in a probability distribution that deviates from the symmetrical normal distribution (bell curve) in a given set of data.

Kurtosis refers to the degree of presence of outliers in the distribution.

3.11 Test your knowledge

1. Explain the objectives of Measuring Dispersion
2. What are the characteristics of Good Measure of Dispersion
3. What are different types of Measures of Dispersion
4. What is Range
5. Explain Inter Quartile Range
6. What is Mean Deviation
7. What is Standard Deviation
8. Explain Lorenz Curve
9. **Find the quartiles of the following data: 4, 6, 7, 8, 10, 23, 34.**

3.12 Further Readings

1. Sankalp Gaurav, Business Statistics, Agra Book International
2. Alhance D.N. and Alhance Beena, Fundamental of Statistics.
3. Monga, G.S., Elementary Statistics.
4. Gupta, S.B., Principles of Statistics.

BLOCK II PROBABILITY

UNIT 4 PROBABILITY – 1

UNIT STRUCTURE:

4.0 OBJECTIVES: After going through this unit you should be able to know about the–

1. Meaning and Definition of Probability.
2. different terminology used in Probability
3. Types of events in Probability

4.1 INTRODUCTION

Probability is a branch of mathematics that deals with measuring the likelihood of an event occurring. It's a fundamental concept in statistics, engineering, economics, and many other fields. In this unit we will explore the terminology used in probability.

4.2 EXPERIMENT

In probability, an experiment is a process that produces a set of outcomes. An experiment can be:

1. Randomized: Outcomes are uncertain and unpredictable (e.g., coin toss).

2. Deterministic: Outcomes are fixed and predictable (e.g., rolling a die with a fixed outcome).

4.3 SAMPLE SPACE (S)

The sample space is the set of all possible outcomes of an experiment. It's denoted by S and consists of:

1. Individual outcomes (Ω)

2. All possible combinations of outcomes

Types of Sample Spaces:

1. Discrete Sample Space: Finite or countable outcomes (e.g., coin toss, rolling a die).

2. Continuous Sample Space: Uncountable outcomes (e.g., measuring height, time).

4.4 EVENT

In probability, an event is a subset of the sample space. It's a set of one or more outcomes.

Key Concepts:

1. Simple Event: A single outcome (e.g., rolling a 6).

2. Compound Event: Multiple outcomes (e.g., rolling an even number).

3. Mutually Exclusive Events: Cannot occur together (e.g., rolling a 6 and a 3).

4. Independent Events: Occurrence of one doesn't affect the other (e.g., two separate coin tosses).

4.5 TYPES OF EVENTS IN PROBABILITY

Even though there can be only one sample space for a particular experiment, the space can give rise to different types of events. Following are some types of events in probability:

4.5.1 Impossible and sure events: Any event that can never happen is known as an impossible event. Since such an event will not occur, its probability is always 0. An example of such an event would be sound travelling in a vacuum or zero dividing a number and the quotient being a real number. On the other hand, a sure event is one that will definitely happen. The probability of such an event will always be 1. An example of such an event is the square of 2 is 4 or the moon revolving around the earth.

4.5.2 Complementary events: When the probability of one event is one only and only if the probability of another event is zero, then the two are called complementary events. Simply put, when there are two events such that one can occur only if the other does not, the two events are called complementary to each other. If the probabilities of a complementary pair of events are added, the result is always 1.

An example of a complementary pair of events would be a coin landing on either heads or tails. One event can occur only when the other does not.

4.5.3 Exhaustive events: The set of all possible events from a sample space from which one event will always occur is called exhaustive events. This means that all the outcomes for an experiment are exhaustive events because at least one of them will definitely happen. The heads or tails of a coin are exhaustive events since at least one of them definitely happens.

Another example would be the die landing on a particular number. The outcome will be one of 1,2,3,4,5 or 6. So all these outcomes form the exhaustive events of rolling a die since one of them will compulsorily occur. So the sum of exhaustive events makes up the whole of the sample space. But this does not mean that all the events in the set of exhaustive events are equally likely to happen. Not all events in the set of exhaustive events share the same probability of occurrence.

4.5.4 Equally likely events: When the probability of events is equal, they are said to be equally likely events. For example, it is equally likely that the coin will land on heads as it is that it will land on tails. So the two events are equally likely events.

4.5.5 Mutually exclusive events: Suppose there are two events such that one cannot occur if the other does, then the two events are said to be mutually exclusive. For example, suppose there is a sample space set S = {13,14,15,16,17,18,19} and there are two subsets A = {13,15,17} and B = {14,16,18}. The two subsets A and B do not have anything in common, so they are mutually exclusive.

4.5.6 Simple and compound events: A result is a simple event if it is a single point from the sample space. For example, from a sample of 1,2,3,4,5,6, the occurrence of a result that is less than 2 can only be one and can be represented by the single number 1. If the event encompasses more than one result from the sample, then the event is said to be compound. For instance, the occurrence of a result that is less than 5 is a compound event since it can be anything from 1 to 4.

4.5.7 Independent and dependent events: When an event does not depend on the outcome of a previous event, then it is known as an independent result. The probability of the occurrence of an independent event will always be the same irrespective of the number of times the experiment has been carried out. The occurrences of a coin landing on heads or tails are independent events. On the other hand, dependent events are affected by the occurrence of some previous event.

4.6 PROBABILITY MEASURE

A probability measure is a function that assigns a number between 0 and 1 to each event in a sample space, representing its likelihood.

4.7 PROBABILITY RULES:

1. Axioms:
   - The probability of an event is between 0 and 1.
   - The probability of the sample space is 1.
   - The probability of the empty set is 0.
2. Addition Rule: P(A or B) = P(A) + P(B) - P(A and B)
3. Multiplication Rule: P(A and B) = P(A) × P(B) (if A and B are independent)

4.8 SUMMARY

Probability is a branch of mathematics that deals with measuring the likelihood of an event occurring. It's a fundamental concept in statistics, engineering, economics, and many other fields

In probability, an experiment is a process that produces a set of outcomes.

The sample space is the set of all possible outcomes of an experiment. It's denoted by S

In probability, an event is a subset of the sample space. It's a set of one or more outcomes.

Impossible and sure events: Any event that can never happen is known as an impossible event. Since such an event will not occur, its probability is always 0.

Complementary events: When the probability of one event is one only and only if the probability of another event is zero, then the two are called complementary events.

Exhaustive events: The set of all possible events from a sample space from which one event will always occur is called exhaustive events.

Equally likely events: When the probability of events is equal, they are said to be equally likely events.

Mutually exclusive events: Suppose there are two events such that one cannot occur if the other does, then the two events are said to be mutually exclusive.

Mutually exclusive events: Suppose there are two events such that one cannot occur if the other does, then the two events are said to be mutually exclusive.

Simple and compound events: A result is a simple event if it is a single point from the sample space

Independent and dependent events: When an event does not depend on the outcome of a previous event, then it is known as an independent result. The probability of the occurrence of an independent event will always be the same irrespective of the number of times the experiment has been carried out.

4.9 Test Your Knowledge

1. What is probability?
2. Explain the various types of Probability Events.
3. What are equally likely events
4. Write short notes on

a. Exhaustive events

b. Sample space

c. Independent event

d. Mutually exclusive event

e. Rules of portability

4.10 Further study

1. Sankalp Gaurav, Business Statistics, Agra Book International
2. Alhance D.N. and Alhance Beena, Fundamental of Statistics.
3. Monga, G.S., Elementary Statistics.
4. Gupta, S.B., Principles of Statistics.

UNIT – 5 PROBABILITY 2

UNIT STRUCTURE:

1. Meaning and Definition of Probability.
2. Probability Defined.
3. Probability Theorems
4. Addition Theorem
5. Multiplication Theorem
6. Summary
7. Test Your Knowledge
8. Further Study

OBJECTIVES:

After going through this unit you should be able to know about the–

1. Meaning and Definition of Probability.
2. Theorems of Probability
3. Calculation of probability by applying law of addition and multiplication

5.1 MEANING AND DEFINITION OF PROBABILITY

One of the Primary reasons for the development of the Theory of Probability is the presence in almost every aspect of life, of random phenomena. A Phenomenon is random if change factors — determine its outcome. All the possible outcomes may be known in advance, but the particular outcome of a single trial in any experimental operation can't be pre-determined. Nevertheless, some regulatory is built into the process so that each of the possible outcomes can be assigned a probability fraction. Probability is especially important in statistics because of the many principles and Procedures that are based on this concept. Indeed Probability plays a special role in all our lives, and has an everyday meaning. Sometimes we hear phases like : 'You had better take an umbrella because it is likely to rain.' 'His changes of winning are preety small'. It is very likely that it will rain by the evening.' 'You are probably right.' Or 'There are fifty-fifty changes of his passing the examinations.' In each of these phrases an idea of uncertainty is acknowledged. Goeth remarked that, "There is nothing more frightful than action in ignorance."

5.2 PROBABILITY DEFINED

Ordinarily speaking, the probability of an event denotes the likelihood of its happening. The value of probability ranges between 0 and 1. If an event is certain to happen, its probability would be 1 and if it is certain that the event would not take place, then the probability of its happening is 0. Ordinarily, in social sciences, the probability of the happening of an event is rarely 1 or 0. The reason is that in social sciences we deal with situations where there is always an element of uncertainty about the happening or not happening of an event. For this reason, the probability of the events is somewhere between 0 and 1.

The general rule of the happening of an event is that if an event can happen in m ways and fail to happen in n ways, then the probability (p) of the happening of the event is given by :

$$p = \frac{m}{m+n}$$

or $p = \frac{\text{Number of cases favourable to the occurrence of the event, i.e., m}}{\text{Total number of mutually exclusive, equally likely and exhaustive cases, i.e., (m + n)}}$

If m = 0, p = 0

If n = 0, p = 0

If m = 0, there is no case favourable to the occurrence of the event and the event is said to be impossible and if n = 0, there is no case unfavourable to the occurrence of the event and the event is said to be sure or certain.

odds in favour of the occurrence of the event

$$= \frac{n}{m} = \frac{\text{The number of cases favourable to the occurrence of the event}}{\text{The number of cases against the occurrence of the event}}$$

odds against the occurrence of the event

$$= \frac{m}{n} = \frac{\text{The number of cases against the occurrence of the event}}{\text{The number of cases favourable to the occurrence of the event}}$$

Notes : $P(A) + P(\overline{A}) = 1 \Rightarrow P(A) = 1 - P(\overline{A})$

i.e., the probability of the occurrence of an event, say, A.

= I – Probability of the occurrence of the event complementary to A.

e.g., if a universal set U = {1, 2, 3, 4, 5, 6, 7, 8, 9} and Set A = {1, 5, 6},

then $\overline{A}$ = {2, 3, 4, 7, 8, 9}

Probability of the occurrence of any event in the sample points of A

$$= P(A) = \frac{3}{9} = \frac{1}{3}, \text{ and } P(\overline{A}) = 1 - \frac{1}{3} = \frac{2}{3}$$

(2) If the events are mutually exclusive and exhaustive, i.e., if the events are complementary the sum of their individual probabilities = 1.

(3) Odds in favour of the occurrence of the event and the odds against the occurrence are reciprocal of each other.

5.3 PROBABILITY THEOREMS

The solution to many problems involving. Probabilities requires a through understanding of some of basic rule which govern the manipulation of probabilities. They are generally called probability theorems. They are discussed below :

5.4 ADDITION THEOREM:

The theorem is stated as follows: "It two events are mutually exclusive and the Probability of the one is P_1 while that of other is P_2, the probability of either the one events or the other occurring is the sum $P_1 + P_2$".

For example, the Probability of getting spot (1) in a throw of a single die is 1/6 the Probability of getting spot

(3) is also 1/6 and the Probability of getting spot (5) too is 1/6. The Probability of getting an old number (1, 3, 5) in a throw of a single die will be the addition of their respective Probabilities, that is —

$$\frac{1}{6}+\frac{1}{6}+\frac{1}{6}=\frac{3}{6} \quad \text{or} \quad \frac{1}{2}$$

The addition theorem will hold good only if:

(i) Item are mutually exclusive,

(ii) Mutually exclusive items belong to same set.

Illustration

(a) A bag contains 4 white, 2 black, 3 yellow and 3 red balls. What is the Probability of getting a white or red ball at random in a single draw of one.

The Probability of getting one white balls = $\frac{4}{12}$

The Probability of getting one red ball = $\frac{3}{12}$

The Probability of getting one white or red ball

$$=\frac{4}{12}+\frac{3}{12}=\frac{7}{12} \quad \text{or} \quad \frac{7}{12}\times 100=58.3\%$$

= 58.3% **Ans.**

(b) Find out the Probability of getting a total of either 7 or 11 in a single throw with two dice.

A total of 7 cm come in 6 different ways —

$$\begin{pmatrix}1 & 2 & 3 & 4 & 5 & 6\\ 6 & 5 & 4 & 3 & 2 & 1\end{pmatrix}$$

A total of 11 can come 2 different ways —

$$\begin{pmatrix}5 & 6\\ 6 & 5\end{pmatrix}$$

The Probability of getting a total of $7=\frac{6}{36}=\frac{1}{6}$

The Probability of getting a total of $11=\frac{2}{36}=\frac{1}{18}$

The Probability of getting either 7 or 11 =

$$\frac{1}{6}+\frac{1}{18}=\frac{4}{18}=\frac{2}{9}$$

The addition theorem will hold good only if the events are mutually exclusive. If events contain no sample point in common, them some adjustment is necessary. Under such a case —

$$P[(A) \text{ or } (B)] = P(A) + P(B) - P(A \& B)$$

The following example will make it clear –

Example – A bag contain 25 balls, numbered from 1 to 25, one is to be drawn at random. Find the Probability that the number of the drawn ball will be multiple of 5 or of 7.

The Probability of number being multiple of

$$5\ (5, 10, 15, 20, 25) = \frac{5}{25} \text{ or } \frac{1}{5}$$

The Probability of the number being multiple of

$$7\ (7, 14, 21) = \frac{3}{25}$$

Thus the Probability of the number being a multiple of 5 or 7 will be –

$$\frac{5}{25} + \frac{3}{25} = \frac{8}{25}$$

in the above illustration, find the Probability that the number is multiple of 3 or 5 :

The Probability of the number being multiple of

$$3\ (3, 6, 9, 12, 15, 18, 21, 24) = \frac{8}{25}$$

The Probability of the number being multiple of

$$5\ (5, 10, 15, 20, 25) = \frac{5}{25}$$

Joint Probability $\frac{8}{25}+\frac{5}{25}=\frac{13}{25}$, but this answer is wrong, because item no. 15 is not mutually exclusive. Hence the correct Probability will be –

$$\frac{8}{25}+\frac{5}{25}-\frac{1}{25}=\frac{12}{25}$$

Hence, $P(A + B) = P(A) + P(B) - P(AB)$.

Example – A card is drawn at random from an ordinary Pack of 52 playing cards. Find the Probability that a card drawn is either a spade or the ace of diamonds.

The Probability of drawing a spade = $\frac{13}{12}$.

The Probability of drawing and ace of diamonds = $\frac{1}{52}$.

Probability of drawing a spade on an ace of diamonds =

$$\frac{13}{52}+\frac{1}{52}=\frac{14}{52} \text{ or } \frac{7}{26}$$ **Ans.**

5.5 MULTIPLICATION THEOREM

According to this theorem, "If two events are mutually independent, and the Probability of the one is P_1 while that of the other is P_2, the Probability of the two events occurring simultaneously is the product of P_1 and P_2." For example, the Probability of head coming up in a toss of a coin is $\frac{1}{2}$ and the Probability of 4 coming in a throw of a die is $\frac{1}{6}$, if a coin and a die are thrown together, the Probability of head coming up in the toss of the coin and 4 coming up in the throw of a die will be $\frac{1}{2}\times\frac{1}{6}=\frac{1}{12}$.

Illustration

(a) What is the Probability of throwing two 'fours' in two throws of a die ?

The Probability of a 'four' in first throw = $\frac{1}{6}$.

The Probability of a four in second throw = $\frac{1}{6}$.

The Possibility of two 'fours' = $\frac{1}{6}\times\frac{1}{6}=\frac{1}{36}$ **Ans.**

(b) What is the Probability of getting all the heads in four throws of a coin ?

The change of getting head in the 1st throw = $\frac{1}{2}$.

The change of getting head in the 2nd throw = $\frac{1}{2}$.

The change of getting head in the 3rd throw = $\frac{1}{2}$.

The change of getting head in the 4th throw = $\frac{1}{2}$.

Thus the Probability of getting heads in all the throws :

$$\frac{1}{2}\times\frac{1}{2}\times\frac{1}{2}\times\frac{1}{2}=\frac{1}{16}$$ **Ans.**

(c) A Problem in statistics is given to three students A, B, C whose changes of solving it are 1/2, 1/3, 1/4 respectively. What is the Probability that the Problem will be solved ?

Probability that student A will fail to solve the Problem = $1-\frac{1}{2}=\frac{1}{2}$.

Probability that student B will fail to solve the Problem = $1-\frac{1}{3}=\frac{2}{3}$.

Probability that student C will fail to solve the Problem = $1-\frac{1}{4}=\frac{3}{4}$.

Since the events are independent, the Probability that all the students A, B, C will fail to solve the Problem –

$$\frac{1}{2}\times\frac{2}{3}\times\frac{3}{4}=\frac{1}{4}.$$

$\therefore$ The Probability that the Problem will be solved = $1-\frac{1}{4}=\frac{3}{4}$.

This Problem can also be solved in the following way :

Condition	**Probability**
(i) A Solve, B Solve, C Solve	= $\frac{1}{2}\times\frac{1}{3}\times\frac{1}{4}=\frac{1}{24}$
(ii)A Solve, B Solve, C fails to solves	= $\frac{1}{2}\times\frac{1}{3}\times\frac{3}{4}=\frac{3}{24}$
(iii) A solve, B fails to solve, C solve = $\frac{1}{2}\times\frac{2}{3}\times\frac{1}{4}=\frac{2}{24}$	
(iv) A fails to solve, B solve, C solve = $\frac{1}{2}\times\frac{1}{3}\times\frac{1}{4}=\frac{1}{24}$	
(v)A solve, B fails to solve, C fails to solve	= $\frac{1}{2}\times\frac{2}{3}\times\frac{3}{4}=\frac{6}{24}$
(vi) A fails to solve, B solve, C fails to solve = $\frac{1}{2}\times\frac{1}{3}\times\frac{3}{4}=\frac{3}{24}$	
(vii) A fails to solve, B fails to solve, C solve = $\frac{1}{2}\times\frac{2}{3}\times\frac{1}{4}=\frac{2}{24}$	
(viii) A, B, C fails to solve	= $\frac{1}{2}\times\frac{2}{3}\times\frac{3}{4}=\frac{6}{24}$

The Problem is solved in all the conditions, except that of (viii). If the Probabilities of (i) to (vii) are added, that will give the Probability of Problem being solved. The total comes to $\frac{18}{24}$ or $\frac{3}{4}$.

The multiplication theorem will hold good only if the events belong to the same set. In order to show the importance of this fact. Moroney in his book "Facts from Figures" given an interesting example. His observe, "Consider the case of a man who demands the simultaneous occurrence of money virtue of an unrelated nature in his young lady. Let's suppose that he insists on a Grecian nose. Plantinum-blonde hair, eyes of odd colours one blue, one brown, and finally a first class knowledge of statistics. What is the Probability that the first lady he needs in the street will put ideas of marriage into his head? It is difficult to apply multiplication theorem in this case, because events do not belong to the same set.

5.6 SUMMARY

The solution of many problems involving Probabilities requires a through under studying of some of the basic rule which govern manipulation of Probabilities. The basic reason of the development of this theory is the presence in almost every aspect of life of random phenomenon.

5.7 Test Your Knowledge

1. Define Probability? Explain the use of probability in business
2. Explain addition theorem?
3. Explain Multiplication Theorem?

5.8 Further study

1. Sankalp Gaurav, Business Statistics, Agra Book International
2. Alhance D.N. and Alhance Beena, Fundamental of Statistics.
3. Monga, G.S., Elementary Statistics.
4. Gupta, S.B., Principles of Statistics.

UNIT 6 CONDITIONAL THEORY & BAYES THEOREM

UNIT STRUCTURE

6.0 OBJECTIVES

After going through this unit you should be able to know about the–

1. Meaning and Definition of Conditional Theory.
2. Have knowledge of Bayes Theorem
3. Understand the application of Bayes Theorem

6.1 CONDITIONAL THEORY

Conditional Theory is a branch of probability theory that deals with the probability of an event occurring given that another event has occurred. It's a fundamental concept in statistics, machine learning, and data science.

6.2 KEY CONCEPTS OF CONDITIONAL THEORY:

1. Conditional Probability: The probability of an event occurring given that another event has occurred.
2. Conditional Distribution: The probability distribution of a random variable given that another random variable has taken on a specific value.
3. Bayes' Theorem: A mathematical formula that describes how to update the probability of a hypothesis given new evidence.
4. Independence: Two events are independent if the occurrence of one event does not affect the probability of the other event.
5. Conditional Independence: Two events are conditionally independent given a third event if the occurrence of one event does not affect the probability of the other event given the third event.

Rules of Conditional Theory:

1. Chain Rule: $P(A, B) = P(A|B) * P(B)$
2. Product Rule: $P(A|B) = P(A, B) / P(B)$

3. Sum Rule: $P(A|B) = \Sigma P(A, Bi) / P(B)$

6.3 APPLICATIONS OF CONDITIONAL THEORY :

1. Predictive Modeling: Conditional theory is used to build predictive models that take into account the relationships between variables.
2. Inference: Conditional theory is used to make inferences about a population based on a sample of data.
3. Decision Making: Conditional theory is used to make decisions under uncertainty by taking into account the probability of different outcomes.
4. Machine Learning: Conditional theory is used in machine learning algorithms such as Bayesian networks and decision trees.

Here's an example of conditional theory:

Suppose we want to know the probability that a person has a fever (F) given that they have a cold (C).

Let's say:

- $P(F) = 0.2$ (probability of having a fever)

- $P(C) = 0.3$ (probability of having a cold)

- $P(F|C) = 0.6$ (probability of having a fever given that they have a cold)

Using the chain rule, we can calculate:

$P(F, C) = P(F|C) * P(C) = 0.6 * 0.3 = 0.18$

This means that the probability of having both a fever and a cold is 0.18.

Now, let's say we want to know the probability of having a fever given that they have a cold and a headache (H).

Let's say:

- $P(H|F, C) = 0.8$ (probability of having a headache given that they have a fever and a cold)

Using the product rule, we can calculate:

$P(F, H|C) = P(H|F, C) * P(F|C) = 0.8 * 0.6 = 0.48$

This means that the probability of having a fever and a headache given that they have a cold is 0.48.

6.4 BAYES' THEOREM

Bayes' Theorem is a fundamental concept in probability theory and statistics that describes how to update the probability of a hypothesis (H) given new evidence (E). It's named after Thomas Bayes, who first proposed it in the 18th century.

Theorem:

$P(H|E) = P(E|H) * P(H) / P(E)$

Where:

- $P(H|E)$ is the posterior probability of the hypothesis given the evidence

- $P(E|H)$ is the likelihood of the evidence given the hypothesis

- P(H) is the prior probability of the hypothesis

- P(E) is the probability of the evidence

Example:

Suppose we want to determine the probability that a person has a disease (H) given that they test positive (E) for the disease.

- P(H) = 0.01 (prior probability of the disease)

- P(E|H) = 0.9 (likelihood of a positive test given the disease)

- P(E) = 0.1 (probability of a positive test)

Using Bayes' Theorem, we can calculate:

P(H|E) = P(E|H) * P(H) / P(E) = 0.9 * 0.01 / 0.1 = 0.09

So, the probability of the person having the disease given a positive test result is 9%.

6.5 APPLICATION OF BAYES' THEOREM

Bayes' Theorem has numerous applications in various fields, including:

1. Medical Diagnosis: Predicting the probability of a disease given symptoms and test results.
2. Machine Learning: Updating model probabilities based on new data.
3. Data Analysis: Inferring population characteristics from sample data.
4. Decision Making: Making informed decisions under uncertainty.
5. Signal Processing: Filtering out noise and extracting signals.
6. Image Recognition: Classifying images based on features and patterns.
7. Natural Language Processing: Predicting word sequences and sentiment analysis.
8. Finance: Credit risk assessment and portfolio optimization.
9. Marketing: Customer segmentation and targeting.
10. Quality Control: Predicting defect rates and improving manufacturing processes.
11. Genetics: Inferring gene expression and disease susceptibility.
12. Sports Analytics: Predicting game outcomes and player performance.
13. Cyber security: Detecting anomalies and predicting attacks.
14. Environmental Science: Predicting climate patterns and natural disasters.
15. Social Network Analysis: Predicting connections and community structure.

Bayes' Theorem provides a powerful framework for updating probabilities based on new information, making it a fundamental tool in many fields.

Question

A doctor wants to determine the probability that a patient has a rare disease (D) given that they test positive (T) for the disease. The doctor knows that:

- The prevalence of the disease is 2% (P(D) = 0.02)

- The test is 95% accurate for patients with the disease (P(T|D) = 0.95)

- The test is 90% accurate for patients without the disease (P(~T|~D) = 0.90)

Solution:

We want to find P(D|T), the probability of the patient having the disease given a positive test result.

Using Bayes' Theorem:

P(D|T) = P(T|D) * P(D) / P(T)

First, we need to find P(T), the probability of a positive test result.

P(T) = P(T|D) * P(D) + P(T|~D) * P(~D)

= 0.95 * 0.02 + 0.10 * 0.98

= 0.019 + 0.098

= 0.117

Now we can plug in the values:

P(D|T) = P(T|D) * P(D) / P(T)

= 0.95 * 0.02 / 0.117

= 0.162

Result:

The probability of the patient having the disease given a positive test result is approximately 16.2%.

Note:

This result may seem counterintuitive, as the test is 95% accurate for patients with the disease. However, the low prevalence of the disease (2%) means that most positive test results will be false positives.

6.6 SUMMARY

Conditional Theory is a branch of probability theory that deals with the probability of an event occurring given that another event has occurred. It's a fundamental concept in statistics, machine learning, and data science.

1. Conditional Probability: The probability of an event occurring given that another event has occurred.
2. Conditional Distribution: The probability distribution of a random variable given that another random variable has taken on a specific value.
3. Bayes' Theorem: A mathematical formula that describes how to update the probability of a hypothesis given new evidence.
4. Independence: Two events are independent if the occurrence of one event does not affect the probability of the other event.
5. Conditional Independence: Two events are conditionally independent given a third event if the occurrence of one event does not affect the probability of the other event given the third event.

6.7 TEST YOUR PROGRESS

1. What is conditional theory?

.....
.....
.....
.....

2. Explain Bayes theorem

.....
.....
.....
.....

3. With example discuss how conditional probability affects the original probability

.....
.....
.....
.....

6.8 SUGGESTED READINGS

1. Sankalp Gaurav, Business Statistics, Agra Book International
2. Alhance D.N. and Alhance Beena, Fundamental of Statistics.
3. Monga, G.S., Elementary Statistics.
4. Gupta, S.B., Principles of Statistics.

Unit 7 THEORETICAL DISTRIBUTIONS

UNIT STRUCTURE

7.0 OBJECTIVE

After going through this unit you will be able to understand about

- Theoretical Distribution And Its Uses
- Types of Theoretical Distribution
- Benefits of Theoretical Distribution

7.1 INTRODUCTION

Theoretical distributions, also known as probability distributions, are mathematical models that describe the probability of different outcomes or values in a random experiment or phenomenon. They are used to:

1. Model real-world phenomena

2. Analyze and interpret data

3. Make predictions and decisions

7.2 USES OF THEORETICAL DISTRIBUTIONS

Theoretical distributions have numerous practical applications across various fields, including:

1. Statistics and Data Analysis:

   - Hypothesis testing
   - Confidence intervals
   - Regression analysis
   - Time series analysis

2. Engineering:

   - Reliability engineering (failure rates)
   - Quality control (defect rates)

- Signal processing (noise modeling)

3. Finance and Economics:
   - Risk analysis (portfolio optimization)
   - Option pricing (Black-Scholes model)
   - Credit risk modelling
4. Insurance:
   - Actuarial science (claim frequency, severity)
   - Risk assessment
   - Policy pricing
5. Medicine and Healthcare:
   - Clinical trials (patient outcomes)
   - Epidemiology (disease modelling)
   - Medical imaging (image analysis)
6. Computer Science:
   - Machine learning (probability models)
   - Data mining (pattern detection)
   - Network modeling (traffic analysis)
7. Operations Research:
   - Optimization algorithms
   - Simulation modelling
   - Queuing theory
8. Environmental Science:
   - Climate modelling
   - Population dynamics
   - Ecological risk assessment
9. Social Sciences:
   - Survey research (response rates)
   - Political science (election forecasting)
   - Sociology (social network analysis)

Theoretical distributions help:

1. Model complex phenomena

2. Make predictions and forecasts

3. Estimate uncertainty and risk

4. Optimize systems and processes

5. Inform decision-making

6. Analyze and interpret data

7. Identify patterns and trends

8. Develop statistical models

9. Communicate insights effectively

By applying theoretical distributions, professionals can gain valuable insights, make informed decisions, and drive meaningful outcomes in their respective fields.

7.3 TYPES OF THEORETICAL DISTRIBUTIONS

There are several types of theoretical distributions, including:

Discrete Distributions

1. Bernoulli Distribution (binary outcomes)

2. Binomial Distribution (fixed trials, two outcomes)

3. Poisson Distribution (count data, rare events)

4. Geometric Distribution (first success in repeated trials)

5. Negative Binomial Distribution (number of failures before successes)

6. Hypergeometric Distribution (sampling without replacement)

Continuous Distributions

1. Uniform Distribution (equal probabilities)

2. Normal Distribution (Gaussian, symmetric)

3. Exponential Distribution (time between events)

4. Gamma Distribution (continuous, skewed)

5. Chi-Squared Distribution (squared normals)

6. Student's t-Distribution (small samples, continuous)

7. Weibull Distribution (continuous, skewed)

8. Lognormal Distribution (continuous, skewed)

9. Beta Distribution (continuous, bounded)

10. Pareto Distribution (continuous, power-law)

Mixed Distributions

1. Continuous-Binary Mixtures

2. Zero-Inflated Distributions (e.g., zero-inflated Poisson)

Specialized Distributions

1. Multivariate Distributions (multiple variables)

2. Multinomial Distribution (categorical outcomes)

3. Dirichlet Distribution (multivariate, categorical)

4. Wishart Distribution (multivariate, covariance matrices)

5. Inverse Wishart Distribution (multivariate, covariance matrices)

7.4 BENIFITS OF THEORETICAL DISTRIBUTIONS

Theoretical distributions offer numerous benefits, including:

1. **Simplified Modelling**: They provide a mathematical framework to model complex phenomena.

2**. Predictive Power**: Enable predictions and forecasts about future events or outcomes.

3. **Uncertainty Quantification:** Allow estimation of uncertainty and risk.

4. **Data Analysis**: Facilitate data analysis, interpretation, and visualization.

5. **Inference and Decision-Making**: Inform statistical inference, decision-making, and optimization.

6. **Generalizability:** Enable generalization to broader populations or scenarios.

7. **Comparison and Benchmarking:** Facilitate comparison across different groups or scenarios.

8. **Identification of Patterns:** Help identify patterns, trends, and correlations.

9. **Simulation and Scenario Analysis**: Enable simulation and scenario analysis.

10. **Communication**: Facilitate clear communication of complex ideas and results.

11**. Efficiency**: Reduce complexity and increase efficiency in data analysis.

12. **Flexibility:** Can be adapted and combined to model diverse phenomena.

13. **Improved Accuracy:** Provide more accurate models and predictions.

14**. Robustness**: Offer robustness against outliers, noise, and missing data.

15. **Interdisciplinary Applications**: Applicable across various fields and disciplines.

Theoretical distributions provide a powerful toolkit for:

- Statistical analysis
- Data science
- Machine learning

- Engineering
- Economics
- Finance
- Biology
- Medicine
- Social sciences

7.5 SUMMARY

Theoretical distributions, also known as probability distributions, are mathematical models that describe the probability of different outcomes or values in a random experiment or phenomenon. Theoretical distribution is used in the field of science commerce business economics etc. there are three mail type of theoretical distribution Binomial, Poisson and Normal Distribution.

7.6 TEST YOUR KNOWLEDGE

1 Mention the uses of Theoretical Distributions

.....
.....
.....
.....

2 Explain the different types of Theoretical Distributions

.....
.....
.....
.....

3 What are the benefits of Theoretical Distributions?

.....
.....
.....
.....

7.7 FURTHER READINGS

1. Sankalp Gaurav, Business Statistics, Agra book International , Agra
2. Sinha, V.C., Principles of Statistics.
3. Gupta, S.B., Principles of Statistics.
4. Monga, G.S., Elementary Statistics.

UNIT 8 BINOMIAL AND POISON DISTRIBUTION

UNIT STRUCTURE

8.0 OBJECTIVE: After reading this unit you are able to understand

- About binomial and Poisson distribution
- Its use and application

8.1 INTRODUCTION

The binomial distribution is a commonly used discrete distribution in statistics. The normal distribution as opposed to a binomial distribution is a continuous distribution. The binomial distribution represents the probability for 'x' successes of an experiment in 'n' trials, given a success probability 'p' for each trial at the experiment.

8.2 BINOMIAL DISTRIBUTION IN STATISTICS

The binomial distribution forms the base for the famous binomial test of statistical importance. A test that has a single outcome such as success/failure is also called a Bernoulli trial or Bernoulli experiment, and a series of outcomes is called a Bernoulli process. Consider an experiment where each time a question is asked for a yes/no with a series of n experiments. Then in the binomial probability distribution, the boolean-valued outcome the success/yes/true/one is represented with probability p and the failure/no/false/zero with probability q ($q = 1 - p$). In a single experiment when $n = 1$, the binomial distribution is called a Bernoulli distribution.

8.3 WHAT IS THE BINOMIAL DISTRIBUTION FORMULA?

The binomial distribution formula is for any random variable X, given by; $P(x:n,p) = {}^{n}Cx\ p^{x}\ (1-p)^{n-x}$ **Or** $P(x:n,p) = {}^{n}C_{x}\ p^{x}\ (q)^{n-x}$

where,

- n = the number of experiments
- x = 0, 1, 2, 3, 4, …
- p = Probability of success in a single experiment
- q = Probability of failure in a single experiment (= 1 – p)

The binomial distribution formula is also written in the form of n-Bernoulli

trials, where ${}^{n}C_{x} = n!/x!(n-x)!$. Hence, $P(x:n,p) = n!/[x!(n-x)!].p^{x}.(q)^{n-x}$

Example 1: If a coin is tossed 5 times, using binomial distribution find the probability of:

(a) Exactly 2 heads

(b) At least 4 heads.

Solution:

(a) The repeated tossing of the coin is an example of a Bernoulli trial. According to the problem:

Number of trials: n=5

Probability of head: p= 1/2 and hence the probability of tail, q =1/2

For exactly two heads:

x=2

$P(x=2) = {}^{5}C2\ p^{2}\ q^{5-2} = 5!\ /\ 2!\ 3! \times (½)^{2} \times (½)^{3}$

P(x=2) = 5/16

(b) For at least four heads,

$x \geq 4$, $P(x \geq 4) = P(x = 4) + P(x=5)$

Hence,

$P(x = 4) = {}^{5}C4\ p^{4}\ q^{5-4} = 5!/4!\ 1! \times (½)^{4} \times (½)^{1} = 5/32$

$P(x = 5) = {}^{5}C5\ p^{5}\ q^{5-5} = (½)^{5} = 1/32$

Answer: Therefore, $P(x \geq 4) = 5/32 + 1/32 = 6/32 = 3/16$

Example 2: For the same question given above, find the probability of getting at most 2 heads.

Solution:

Solution: P(at most 2 heads) = $P(X \leq 2) = P\ (X = 0) + P\ (X = 1)$

$P(X = 0) = (½)^{5} = 1/32$

$P(X=1) = 5C1\ (½)^5.= 5/32$

Answer: Therefore, $P(X \leq 2) = 1/32 + 5/32 = 3/16$

8.4 ASSUMPTION FOR BINOMIAL DISTRIBUTION

The Binomial Distribution assumes:

1. Fixed number of trials (n): The experiment consists of a fixed number of independent trials.
2. Independent trials: Each trial is independent of the others.
3. Constant probability: The probability of success (p) remains constant across all trials.
4. Binary outcomes: Each trial has only two possible outcomes: success (S) or failure (F).
5. Random sampling: Trials are randomly selected from the population.
6. Large population: The population size is large compared to the sample size.
7. Stationarity: The probability of success remains constant over time.
8. No replacement: Trials are done without replacement (if sampling from a finite population).
9. Identical trials: All trials have the same probability of success.
10. Discrete outcomes: Outcomes are discrete, not continuous.

When these assumptions hold, the Binomial Distribution accurately models the number of successes in n independent trials, with probability p.

8.5 COMMON APPLICATIONS OF BINOMIAL DISTRIBUTION:

- Coin tossing
- Quality control
- Medical trials
- Election forecasting
- Survey research
- Genetics
- Reliability engineering

8.6 POISSON DISTRIBUTION

The Poisson distribution is a discrete probability distribution that models the number of events occurring in a fixed interval of time or space, where:

1. Events occur independently.
2. Events occur at a constant average rate (λ).
3. Events occur randomly.

8.7 KEY CHARACTERISTICS OF POISSON DISTRIBUTION:

1. Discrete: Models count data (0, 1, 2, ...).
2. Parameter λ (lambda): Average rate of events.
3. Mean = Variance = λ.
4. Skewed distribution (right-skewed for $\lambda < 1$, symmetric for $\lambda = 1$).

8.8 PROBABILITY MASS FUNCTION (PMF):

$P(X = k) = (e^{(-\lambda)} * (\lambda^k)) / k!$

where:

X = number of events

k = 0, 1, 2, ...

e = base of the natural logarithm

λ = average rate

8.9 APPLICATIONS OF POISSON DISTRIBUTION:

1. Count data: Number of defects, accidents, or errors.
2. Time-between-events: Time between arrivals, failures, or events.
3. Queuing theory: Modeling waiting times and queue lengths.
4. Biology: Number of organisms, cells, or particles.
5. Finance: Modeling stock prices, trading volume, or credit risk.
6. Quality control: Defect rates, quality metrics.
7. Telecommunications: Call arrivals, network traffic.

8.10 ASSUMPTIONS OF POISSON DISTRIBUTION:

1. Events occur independently.
2. Events occur at a constant average rate (λ).
3. Events occur randomly.
4. Fixed interval of time or space.

8.11 COMMON USES OF POISSON DISTRIBUTION:

1. Modeling rare events.
2. Analyzing count data.
3. Estimating event rates.
4. Predicting probabilities.

8.12 HOW TO SOLVE PROBLEMS OF POISSON DISTRIBUTION?

To solve problems involving the Poisson Distribution, follow these steps:

1. Identify the problem type:
   - Probability calculation
   - Expected value calculation
   - Finding λ (average rate)
2. Understand the given information:
   - Average rate (λ)
   - Time interval (t)
   - Number of events (k)
3. Choose the appropriate Poisson formula:
   - Probability: P(X = k) = (e^(-λ) * (λ^k)) / k!
   - Expected value: E(X) = λ
   - Variance: Var(X) = λ
4. Plug in the values:
   - Replace λ, k, and t with the given values
5. Calculate the result:
   - Use a calculator or software (e.g., R, Python, Excel) to compute the value
6. Interpret the result:
   - Understand the meaning of the calculated probability, expected value, or variance

Example Problems:

1. Probability:
   - Average defects per unit: 2%
   - Units produced: 1000
   - Find P(X = 3)
   - Solution: P(X = 3) = (e^(-2) * (2^3)) / 3! ≈ 0.180
2. Expected Value:
   - Average calls per minute: 5
   - Time interval: 5 minutes
   - Find E(X)
   - Solution: E(X) = λ * t = 5 * 5 = 25

3. Finding λ:

- Number of accidents per week: 3

- Time interval: 1 week

- Find λ

- Solution: λ = 3 (average rate per week)

Tips:

- Ensure correct units for λ and t

- Use software or calculators for complex calculations

- Interpret results in context

- Verify assumptions (independence, constant rate)

Common mistakes:

- Incorrect λ or t values

- Forgetting to square λ in variance calculation

- Misinterpreting results

Practice problems:

1. A hospital receives an average of 8 patients per hour. Find the probability of exactly 5 patients in the next hour.

2. A manufacturing process produces an average of 2 defects per 100 units. Find the expected number of defects in 500 units.

3. A call centre receives an average of 12 calls per minute. Find the variance of calls in 5 minutes.

8.13 SUMMARY

The binomial distribution is a commonly used discrete distribution in statistics. The normal distribution as opposed to a binomial distribution is a continuous distribution. The binomial distribution represents the probability for 'x' successes of an experiment in 'n' trials, given a success probability 'p' for each trial at the experiment.

The binomial distribution forms the base for the famous binomial test of statistical importance. A test that has a single outcome such as success/failure is also called a Bernoulli trial or Bernoulli experiment, and a series of outcomes is called a Bernoulli process. Consider an experiment where each time a question is asked for a yes/no with a series of n experiments. Then in the binomial probability distribution, the boolean-valued outcome the success/yes/true/one is represented with probability p and the failure/no/false/zero with probability q ($q = 1 - p$). In a single experiment when $n = 1$, the binomial distribution is called a Bernoulli distribution.

The Poisson distribution is a discrete probability distribution that models the number of events occurring in a fixed interval of time or space

8.14 TEST YOUR KNOWLEDGE

1. What is binomial distribution?
2. What is Poisson distribution?
3. Explain Binomial Distribution in Statistics
4. What Is The Binomial Distribution Formula?
5. Mention the assumption for Binomial Distribution
6. Elaborate some Common Applications of Binomial Distribution
7. Define Poisson Distribution?
8. What are the key characteristics of Poisson distribution?
9. What is Probability Mass Function (Pmf)?
10. Explain the applications of Poisson distribution
11. What are the assumptions of Poisson distribution?
12. State some common Uses of Poisson distribution?
13. How to Solve Problems of Poisson distribution?

8.15 FURTHER READINGS

1. Sankalp Gaurav, Business Statistics, Agra book International , Agra
2. Sinha, V.C., Principles of Statistics.
3. Gupta, S.B., Principles of Statistics.
4. Monga, G.S., Elementary Statistics.

UNIT 9 NORMAL DISTRIBUTION

UNIT STRUCTURE

9.0 OBJECTIVE: After going through this unit you are able to understand about

- Normal distribution and its properties
- Application of normal distribution
- Types of normal distribution

9.1 INTRODUCTION

The Normal Distribution, also known as the Gaussian Distribution, is a continuous probability distribution that describes how data points are distributed around a central value, called the mean (μ), with a specific amount of variation, measured by the standard deviation (σ).

9.2 KEY CHARACTERISTICS OF NORMAL DISTRIBUTION

The key characteristics of the Normal Distribution are:

1. Symmetric Bell-Shaped Curve: The normal distribution is symmetric around the mean, with the majority of data points clustered around the center.

2. Mean (μ) = Median = Mode: The mean, median, and mode are equal in a normal distribution.

3. Standard Deviation (σ): The standard deviation measures the spread or dispersion of the data.

4. Continuous Distribution: The normal distribution is a continuous probability distribution.

5. Infinite Range ($-\infty$ to ∞): The normal distribution has an infinite range, extending from negative infinity to positive infinity.

6. Probabilities Calculated Using Z-Score Formula: Probabilities are calculated using the z-score formula: $z = (X - \mu) / \sigma$.

7. Skewness: 0 (Symmetric): The normal distribution is symmetric, with a skewness of 0.

8. Kurtosis: 3 (Mesokurtic): The normal distribution has a kurtosis of 3, indicating a mesokurtic distribution.

9. Asymptotic: The normal distribution approaches the x-axis asymptotically.

10. Area Under the Curve: The total area under the normal distribution curve is 1.

11. Percentages:

- 68% of data points fall within 1 standard deviation ($\mu \pm \sigma$)
- 95% of data points fall within 2 standard deviations ($\mu \pm 2\sigma$)
- 99.7% of data points fall within 3 standard deviations ($\mu \pm 3\sigma$)

9.3 PROPERTIES OF NORMAL DISTRIBUTION

The Normal Distribution has the following properties:

1. Mean (μ): The average value of the distribution.
2. Standard Deviation (σ): Measures the spread or dispersion.
3. Variance (σ^2): The square of the standard deviation.
4. Symmetry: The distribution is symmetric around the mean.
5. Bell-Shaped: The distribution has a bell-shaped curve.
6. Continuous: The distribution is continuous, with no gaps or jumps.
7. Infinite Range: The distribution extends from negative infinity to positive infinity.
8. Probabilistic: The area under the curve represents probability.
9. Total Area = 1: The total area under the curve is equal to 1.
10. Skewness = 0: The distribution is symmetric, with zero skewness.
11. Kurtosis = 3: The distribution has a kurtosis of 3, indicating a mesokurtic distribution.
12. Additivity: The sum of normally distributed variables is also normally distributed.
13. Scalability: The distribution can be scaled by multiplying by a constant.
14. Stability: The distribution is stable under convolution.
15. Maximum Entropy: The normal distribution has maximum entropy among continuous distributions.
16. Central Limit Theorem: The distribution of sample means approaches normality.
17. Invariance: The distribution is invariant under linear transformations.
18. Universality: The normal distribution appears in many natural phenomena.
19. Differentiability: The distribution is infinitely differentiable.
20. Analytical Tractability: The distribution has a closed-form expression.

9.4 APPLICATION OF NORMAL DISTRIBUTION

The Normal Distribution has numerous applications in various fields, including:

1. Statistics: Hypothesis testing, confidence intervals, regression analysis.

2. Science:
   - Physics: Measurement errors, particle distributions.
   - Biology: Population growth, genetic variation.
   - Chemistry: Chemical reactions, concentration measurements.
3. Engineering:
   - Quality control: Manufacturing tolerances, defect rates.
   - Signal processing: Noise analysis, filter design.
   - Reliability engineering: Failure rates, maintenance scheduling.
4. Finance:
   - Stock prices: Modeling returns, risk analysis.
   - Portfolio management: Asset allocation, risk assessment.
   - Option pricing: Black-Scholes model.
5. Medicine:
   - Clinical trials: Patient outcomes, treatment efficacy.
   - Epidemiology: Disease spread, risk factors.
   - Medical research: Data analysis, hypothesis testing.
6. Social Sciences:
   - Psychology: IQ scores, personality traits.
   - Sociology: Population demographics, social phenomena.
   - Economics: Income distribution, economic indicators.
7. Business:
   - Marketing: Customer behavior, market research.
   - Operations research: Supply chain management, inventory control.
   - Management: Decision-making, risk assessment.
8. Environmental Studies:
   - Climate modeling: Temperature distributions, weather patterns.
   - Ecology: Population dynamics, species distribution.
9. Computer Science:
   - Machine learning: Data modeling, algorithm development.
   - Data analysis: Data visualization, statistical inference.
10. Quality Control:

- Manufacturing: Process control, defect detection.
- Service industry: Quality assessment, customer satisfaction.

9.5 TYPES OF NORMAL DISTRIBUTION

There are several types of Normal Distributions:

1. Standard Normal Distribution (Z-Distribution):
   - Mean (μ) = 0
   - Standard Deviation (σ) = 1
   - Used for standardized scores and probabilities
2. Non-Standard Normal Distribution:
   - Mean (μ) $\neq$ 0
   - Standard Deviation (σ) $\neq$ 1
   - Used for real-world data with varying means and standard deviations
3. Univariate Normal Distribution:
   - Single variable or attribute
   - Used for analyzing individual characteristics
4. Multivariate Normal Distribution:
   - Multiple variables or attributes
   - Used for analyzing relationships between variables
5. Bivariate Normal Distribution:
   - Two variables or attributes
   - Used for analyzing relationships between two variables
6. Truncated Normal Distribution:
   - Limited range or truncated tails
   - Used for modeling data with boundaries or constraints
7. Censored Normal Distribution:
   - Data is censored or incomplete
   - Used for modeling data with missing or censored values
8. Skewed Normal Distribution:
   - Asymmetric or skewed shape
   - Used for modeling data with non-normal skewness
9. Mixture Normal Distribution:

- Combination of multiple normal distributions

- Used for modeling complex data with multiple sub-populations

10. Bayesian Normal Distribution:

- Incorporates prior knowledge or beliefs

- Used for Bayesian inference and modeling

9.6 SUMMARY

The Normal Distribution, also known as the Gaussian Distribution, is a continuous probability distribution that describes how data points are distributed around a central value, called the mean (μ), with a specific amount of variation, measured by the standard deviation (σ).

9.7 TEST YOUR KNOWLEDGE

1. Explain the key characteristics of normal distribution
2. What are the properties of normal distribution?
3. Explain the application of normal distribution
4. What are the different types of normal distribution?

9.8 FURTHER READINGS

1. Sankalp Gaurav, Business Statistics, Agra book International , Agra
2. Sinha, V.C., Principles of Statistics.
3. Gupta, S.B., Principles of Statistics.
4. Monga, G.S., Elementary Statistics.

BLOCK III SAMPLING

UNIT-10: SAMPLING

UNIT FRAMEWORK

10.1 OBJECTIVE

After reading this unit you will be able to:

- The concept of sampling.
- The merits and demerits of sampling.

10.2 INTRODUCTION: MEANING OF SAMPLING

Data collection stage of any research requires considerable time, effort, and money. If primary data are collected using census method, time and cost increases considerably. Sampling techniques help us in this situation. A true representative sample not only gives accurate results but also saves on time, effort, and money. This chapter is devoted to sampling methods and techniques.

The terminology "sampling" indicates the selection of a part of a group or an aggregate with a view to obtaining information about the whole. This aggregate or the totality of all members is known as Population although they need not be human beings. The selected part, which is used to ascertain the characteristics of the population, is called Sample. While choosing a sample, the population is assumed to be composed of individual units or members, some of which are included in the sample. The total number of members of the population is called Population Size and the number included in the sample is called Sample Size.

Researchers usually cannot make direct observations of every individual in the population they are studying. Instead, they collect data from a subset of individuals – a *sample* – and use those observations to make inferences about the entire population.

Ideally, the sample corresponds to the larger population on the characteristic(s) of interest. In that case, the researcher's conclusions from the sample are probably applicable to the entire population.

This type of correspondence between the sample and the larger population is most important when a researcher wants to know what proportion of the population has a certain characteristic –like a particular opinion or a demographic feature. Public opinion polls that try to describe the percentage of the population that plans to vote for a particular candidate, for example, require a sample that is highly representative of the population.

10.3 NEED OF SAMPLING

To draw conclusions about populations from samples, we must use inferential statistics which enables us to determine a population`s characteristics by directly observing only a portion (or sample) of the population. We obtain a sample rather than a complete enumeration (a census) of the population for many reasons. Obviously, it is cheaper to observe a part rather than the whole, but we should prepare ourselves to cope with the dangers of using samples. In this tutorial, we will investigate various kinds of sampling procedures. Some are better than others but all may yield samples that are inaccurate and unreliable. We will learn how to minimize these dangers, but some potential error is the price we must pay for the convenience and savings the samples provide.

ESSENTIALS OF SAMPLING:

In order to reach a clear conclusion, the sampling should possess the following essentials:

1. It must be representative: The sample selected should possess the similar characteristics of the original universe from which it has been drawn.

2. Homogeneity: Selected samples from the universe should have similar nature and should not have any difference when compared with the universe.

3. Adequate samples: In order to have a more reliable and representative result, a good number of items are to be included in the sample.

4. Optimization: All efforts should be made to get maximum results both in terms of cost as well as efficiency. If the size of the sample is larger, there is better efficiency and at the same time the cost is more. A proper size of sample is maintained in order to have optimized results in terms of cost and efficiency.

10.4 ADVANTAGES OF SAMPLING

The sampling only chooses a part of the units from the population for the same study. The sampling has a number of advantages as compared to complete enumeration due to a variety of reasons. Sampling has the following advantages:

1. **Cost effective:** This method is cheaper than the Census Research because only a fraction of the population is studied in this method.
2. **Time saving:** There is saving in time not only in conducting the sampling enquiry but also in the decision making process
3. **Testing of Accuracy:** Testing of accuracy of samples drawn can be made by comparing two or more samples.
4. **Detailed Research is Possible:** Since the data collected under this method is limited but homogeneous, so more time could be spend on decision making.
5. **Reliability:** If samples are taken in proper size and on proper grounds the results of sampling will be almost the same which might have been obtained by Census method.
6. **Exclusive methods in many circumstances**: Where the population is infinite, then the sampling method is the only method of effective research. Also, if the population is perishable or testing units are destructive, then we have to complete our research only through sampling. Example: Estimation of expiry dates of medicines.
7. **Administrative convenience:** The organization and administration of sample survey are easy for the reasons which have been discussed earlier.

8. **More scientific:** Since the methods used to collect data are based on scientific theory and results obtained can be tested, sampling is a more scientific method of collecting data.

10.5 LIMITATIONS OF SAMPLING

It is not that sampling is free from demerits or shortcomings. There are certain limitations of this method which are discussed below:

1. **Biased Conclusion**: If the sample has not been properly taken then the data collected and the decision on such data will lead to wrong conclusion. Samples are like medicines. They can be harmful when they are taken carelessly or without knowledge off their effects.
2. **Experienced Researcher is required**: An efficient sampling requires the services of qualified, skilled and experienced personnel. In the absence of these the results of their search will be biased.
3. **Not suited for Heterogeneous Population**: If the populations are mixed or varied, then this method is not suited for research.
4. **Small Population**: Sampling method is not possible when population size is too small.
5. **Illusory conclusion:** If a sample enquiry is not carefully planned and executed, the conclusions may be inaccurate and misleading.
6. **Sample Not Representative:** To make the sample representative is a difficult task. If a representative sample is taken from the universe, the result is applicable to the whole population. If the sample is not representative of the universe the result may be false and misleading.
7. **Lack of Experts:** As there are lack of experts to plan and conduct a sample survey, its execution and analysis, and its results would be unsatisfactory and not trustworthy.
8. **Conditions of Complete Coverage:** If the information is required for each and every item of the universe, then a complete enumeration survey is better.

10.6 SUMMARY

A statistical sample ideally purports to be a miniature model or replica of the collectivity or the population. Sampling helps in time and cost saving. If the population to be studied is quite large, sampling is warranted. However, the size is a relative matter. The decision regarding census or sampling depends upon the budget of the study. Sampling is opted when the amount of money budgeted is smaller than the anticipated cost of census survey.

10.7 SELF-ASSESSMENT QUESTIONS

1. What do you mean by Sampling? State the purpose of sampling.

   .....
   .....
   .....
   .....

2. Discuss the advantages and limitations of sampling techniques.

   .....
   .....
   .....
   .....

3. “Sampling is necessary under certain conditions”. Explain this with suitable examples.

.....
.....
.....
.....

10.8 TEXT AND REFERENCES

- Donald Cooper, Donald R. Cooper, and Pamela S. Schindler, “Business Research Methods”, Tata McGraw Hill Inc.
- Alan Bryman, and Emma Bell, “Business research Methods”. Oxford University Press Publication.
- Donald S. Tull and Del S. Hawkins, “Marketing Research – Measurement and Method”, Prentice Hall of India Pvt. Ltd., New Delhi.
- William G. Zikmund, “Business Research Methods”, Thomson, South-Western Publication, Singapore.
- C.R. Kothari, “Research Methodology – Methods and Techniques”, Wilely Eastern Limited, Delhi.
- S.P. Gupta, “*Statistical methods*” Sultan Chand & Sons publication, New Delhi
- Saunders, “Research Methods for Business Students”, Pearsons Education Publications.
- Naresh Malhotra, “Marketing Research: An applied Orientation”, Prentice Hall International Edition.
- T.S. Wilkinson, and P.L. Bhandarker, “Methodology and techniques of Social Research”, Himalaya Publishing House, New Delhi.

UNIT 11: SAMPLING METHODS OF DATA COLLECTION

UNIT STRUCTURE

11.0 Objective

After going through this unit you are able to understand

- About data collection
- Methods of data collection and its types
- Primary and secondary data and its difference

11.1 Sampling

Sampling is defined as the process in which the fraction of the population, so selected to represent the characteristics of the larger group. This method is used for statistical testing, where it is not possible to consider all members or observations, as the population size is very large.

As statistical inferences are based on the sampling observations, the selection of the appropriate representative sample is of utmost importance. So, the sample selected should indicate the entire universe and not exhibit a particular section. On the basis of the data collected from the representative samples, the conclusion is drawn from the whole population. **For instance**: A company places an order for raw material by simply checking out the sample.

The units which constitute sample is considered as ‘Sampling Units’. The full-fledged list containing all sampling units is called ‘Sampling Frame’.

11.2 DATA COLLECTION

Data collection is the systematic approach to gathering and measuring information from a variety of sources to get a complete and accurate picture of an area of interest. Data collection enables a person or organization to answer relevant questions, evaluate outcomes and make predictions about future probabilities and trends.

Accurate data collection is essential to maintaining the integrity of research, making informed business decisions and ensuring quality assurance. For example, in retail sales, data might be collected from mobile applications, website visits, loyalty programs and online surveys to learn more about customers. In a server consolidation project, data collection would include not just a physical inventory of all servers, but also an exact description of what is installed on each server -- the operating system, middleware and the application or database that the server supports.

11.3 METHODS FOR COLLECTING STATISTICAL DATA

Let us make an in-depth study of the two methods for collecting statistical data.

11.3.1. Census Method:

The data which are collected by the investigator himself is called primary data. Census data can be thought of as primary data.

When the data collector or investigator collects data or information about each and every item in the population and other related areas, it is known as census method. As this method deals with the investigation of the entire population, it is also called complete enumeration method.

If a survey covers 100 p.c. populations, it is called a census method. In other words, here each and every item or unit constituting the 'entire population' or 'the universe' is selected for statistical enquiry. If a statistical enquiry is conducted to study the nature and pattern of urbanisation, then the universe consists only of the urban population of India. This method is called complete enumeration method because information from each and every unit belonging to India's urban population is collected.

Merits of Census Method:

1. For an extensive study (however expensive it may be) this method is considered to be an ideal one. For example, in the population census, we obtain quite a large number of key information, such as birth rate, death rate, infant mortality rate, literacy rate, ratio of urban-rural population, trend on urbanisation and so on.
2. Acuracy in the results is obtained. The data collected are more accurate and reliable under this census method since information are gathered from various angles. However, reliability in data and their accuracy in results are surely obtained provided enumerators do their work honestly and sincerely.

Demerits of Census Method:

First, being extensive in nature, the complete enumeration method is much expensive since a considerable amount of money, time, and labour are demanded and involved.

Secondly, this method is often not feasible or practicable because the concept of the 'universe' is hypothetical. Since "universe" is the basis of data collection, its applicability becomes very much limited. This method cannot be met with urgency.

This means that if an urgent statistical information for the entire population for policy purposes is needed, this method will surely be less helpful. In other words, it is too cumbersome and inefficient to obtain a complete picture of the target population.

Thirdly, in the census method, often large number of non-sampling errors creeps in. This means that results obtained may not be uniform.

11.3.2. Sample Survey Method:

Instead of the census method, data analysts often consider a portion or a sample of the population. If a sample population—instead of full survey of a population—is investigated then we have sample survey method or portion enumeration method.

A sample is anything less than a full survey of a population. For example, liquor consumption among college and university students is to be investigated. For this purpose, a small number of students following a particular prescribed technique will be picked up (sample drawn) and their habits for consumption of liquor will be investigated. The primary objective of such sample enquiry is to estimate some characteristics of the population from which the sample is selected.

Merits of Survey Method:

Sample survey method has many advantages over census method. That is why this method is more popular than the latter method. In the words of A. C. Rosander, **"If carefully designed, the sample is not only considerably cheaper but may give results which are just accurate and sometimes more accurate than those of a census technique."**

This method is preferable for the following reasons:

1. Since only a part of the population is investigated under this method, it takes less time, less money and less labour. There is saving in time also since sampling enquiry requires less fieldwork, tabulation and data processing than a full survey method. Sample survey method can be conducted when the investigators face the problem of budget constraint. It is cheaper to collect information from a sample group.
2. However, for conducting the entire enquiry, small group of investigators or specialist investigators are paid huge amount and employed but the output is much more. That is why it is also said that this method results in reduced unit cost of enquiry.
3. Conclusions and results obtained from this method are more accurate and reliable as fewer or chosen sample units are surveyed. Trained personnel are usually employed to collect data and investigate the problem. Above all, these people use sophisticated and latest designed techniques so that results become more accurate and reliable. Further, it is true that sampling errors cannot be avoided, but such errors are easier to estimate and control.
4. The small sample data provide a good benchmark for the entire population.
5. From the administrative point of view, the sample method is considered as an ideal one as the organisation as well as administration considers the process practically more convenient. Administrative network does not usually require to be considerably elaborate or extensive.

11.4 TYPES OF DATA

Generally, there are two types of data: quantitative data and qualitative data.

1. Quantitative data is any data that is in numerical form -- e.g., statistics and percentages.
2. Qualitative data is descriptive data -- e.g., colour, smell, appearance and quality.

In addition to quantitative and qualitative data, some organizations might also make use of secondary data to help drive business decisions. Secondary data is typically quantitative in nature and has already been collected by another party for a different purpose. For example, a company might use Indian Census data to make decisions about marketing campaigns. In media, a news team

might use government health statistics or health studies to drive content strategy.

As technology evolves, so does data collection. Recent advancements in mobile technology and the Internet of Things are forcing organizations to think about how to collect, analyze and monetize new data. At the same time, privacy and security issues surrounding data collection heat up.

11.5 PRIMARY AND SECONDARY DATA

In a time when data is becoming easily accessible to researchers all over the world, the practicality of utilizing secondary data for research is becoming more prevalent, same as its questionable authenticity when compared with primary data.

These two types of data, when considered for research is a double-edged sword because it can equally make a research project as well as it can mar it.

In a nutshell, primary data and secondary data both have their advantages and disadvantages. Therefore, when carrying out research, it is left for the researcher to weigh these factors and choose the better one.

It is therefore important for one to study the similarities and differences between these data types so as to make proper decisions when choosing a better data type for research work.

11.6 PRIMARY DATA

Primary data is the kind of data that is collected directly from the data source without going through any existing sources. It is mostly collected specially for a research project and may be shared publicly to be used for other research

Primary data is often reliable, authentic, and objective in as much as it was collected with the purpose of addressing a particular research problem. It is noteworthy that primary data is not commonly collected because of the high cost of implementation.

A common example of primary data is the data collected by organizations during market research, product research, and competitive analysis. This data is collected directly from its original source which in most cases are the existing and potential customers.

Most of the people who collect primary data are government authorized agencies, investigators, research-based private institutions, etc.

Advantages of Primary Data

Following are the advantages of Primary Data:

- Primary data is specific to the needs of the researcher at the moment of data collection. The researcher is able to control the kind of data that is being collected.
- It is accurate compared to secondary data. The data is not subjected to personal bias and as such the authenticity can be trusted.
- The researcher exhibit ownership of the data collected through primary research. He or she may choose to make it available publicly, patent it, or even sell it.
- Primary data is usually up to date because it collects data in real-time and does not collect data from old sources.

- The researcher has full control over the data collected through primary research. He can decide which design, method, and data analysis techniques to be used.
- Disadvantages of Primary Data
- Following are the disadvantages of Primary Data
- Primary data is very expensive compared to secondary data. Therefore, it might be difficult to collect primary data.
- It is time-consuming.
- It may not be feasible to collect primary data in some cases due to its complexity and required commitment.

11.7 SECONDARY DATA

Secondary data is the data that has been collected in the past by someone else but made available for others to use. They are usually once primary data but become secondary when used by a third party.

Secondary data are usually easily accessible to researchers and individuals because they are mostly shared publicly. This, however, means that the data are usually general and not tailored specifically to meet the researcher's needs as primary data does.

For example, when conducting a research thesis, researchers need to consult past works done in this field and add findings to the literature review. Some other things like definitions and theorems are secondary data that are added to the thesis to be properly referenced and cited accordingly.

Some common sources of secondary data include trade publications, government statistics, journals, etc. In most cases, these sources cannot be trusted as authentic.

Advantages of Secondary data

Following are the advantages of Secondary data

- Secondary data is easily accessible compared to primary data. Secondary data is available on different platforms that can be accessed by the researcher.
- Secondary data is very affordable. It requires little to no cost to acquire them because they are sometimes given out for free.
- The time spent on collecting secondary data is usually very little compared to that of primary data.
- Secondary data makes it possible to carry out longitudinal studies without having to wait for a long time to draw conclusions.
- It helps to generate new insights into existing primary data.
- **Disadvantages of Secondary data**
- Following are the Disadvantages of Secondary data
- Secondary data may not be authentic and reliable. A researcher may need to further verify the data collected from the available sources.
- Researchers may have to deal with irrelevant data before finally finding the required data.

- Some of the data is exaggerated due to the personal bias of the data source.
- **Secondary data sources are sometimes outdated with no new data to replace the old ones.**

11.8 SUMMARY

Sampling is defined as the process in which the fraction of the population, so selected to represent the characteristics of the larger group. This method is used for statistical testing, where it is not possible to consider all members or observations, as the population size is very large.

Data collection is the systematic approach to gathering and measuring information from a variety of sources to get a complete and accurate picture of an area of interest. Data collection enables a person or organization to answer relevant questions, evaluate outcomes and make predictions about future probabilities and trends.

Generally, there are two types of data: quantitative data and qualitative data. Quantitative data is any data that is in numerical form -- e.g., statistics and percentages. Qualitative data is descriptive data -- e.g., colour, smell, appearance and quality.

Primary data is the kind of data that is collected directly from the data source without going through any existing sources. It is mostly collected specially for a research project and may be shared publicly to be used for other research

Secondary data is the data that has been collected in the past by someone else but made available for others to use. They are usually once primary data but become secondary when used by a third party.

11.9 TEST YOUR KNOWLEDGE

1. What is sampling?
2. What is data collection?
3. Explain the types of data collection
4. Elaborate the difference between primary and secondary data?
5. What is primary data?

11.10 FURTHER READINGS

- Donald Cooper, Donald R. Cooper, and Pamela S. Schindler, “Business Research Methods”, Tata McGraw Hill Inc.
- Alan Bryman, and Emma Bell, “Business research Methods”. Oxford University Press Publication.
- Donald S. Tull and Del S. Hawkins, “Marketing Research – Measurement and Method”, Prentice Hall of India Pvt. Ltd., New Delhi.
- William G. Zikmund, “Business Research Methods”, Thomson, South-Western Publication, Singapore.
- C.R. Kothari, “Research Methodology – Methods and Techniques”, Wilely Eastern Limited, Delhi.

UNIT 12: SAMPLING DISTRIBUTION

UNIT STRUCTURE

12.1 OBJECTIVE

After reading this unit you will be able to:

- Various probability sampling methods along with their respective merits and demerits.
- Various non-probability sampling methods along with their respective merits and demerits.

12.2 PROBABILITY AND NON-PROBABILITY SAMPLING

A probability sampling is one in which every unit in the population has a chance (greater than zero) of being selected in the sample, and this probability can be accurately determined. The combination of these traits makes it possible to produce unbiased estimates of population totals, by weighting sampled units according to their probability of selection.

Example: We want to estimate the total income of adults living in a given street. We visit each household in that street, identify all adults living there, and randomly select one adult from each household. (For example, we can allocate each person a random number, generated from a uniform distribution between 0 and 1, and select the person with the highest number in each household). We then interview the selected person and find their income. People living on their own are certain to be selected, so we simply add their income to

our estimate of the total. But a person living in a household of two adults has only a one-in-two chance of selection. To reflect this, when we come to such a household, we would count the selected person's income twice towards the total. (The person who is selected from that household can be loosely viewed as also representing the person who isn't selected.) In the above example, not everybody has the same probability of selection; what makes it a probability sample is the fact that each person's probability is known. When every element in the population *does* have the same probability of selection, this is known as an 'equal probability of selection' (EPS) design. Such designs are also referred to as 'self-weighting' because all sampled units are given the same weight.

Probability sampling includes: Simple Random Sampling, Systematic Sampling, and Stratified Sampling, Probability Proportional to Size Sampling, and Cluster or Multistage Sampling. These various ways of probability sampling have two things in common:

1. Every element has a known nonzero probability of being sampled and

2. Involves random selection at some point.

Non-probability sampling is any sampling method where some elements of the population have *no* chance of selection (these are sometimes referred to as 'out of coverage'/'under covered'), or where the probability of selection can't be accurately determined. It involves the selection of elements based on assumptions regarding the population of interest, which forms the criteria for selection. Hence, because the selection of elements is non random, non-probability sampling does not allow the estimation of sampling errors. These conditions give rise to exclusion bias, placing limits on how much information a sample can provide about the population. Information about the relationship between sample and population is limited, making it difficult to extrapolate from the sample to the population.

Example: We visit every household in a given street, and interview the first person to answer the door. In any household with more than one occupant, this is a non-probability sample, because some people are more likely to answer the door (e.g. an unemployed person who spends most of their time at home is more likely to answer than an employed housemate who might be at work when the interviewer calls) and it's not practical to calculate these probabilities.

Non-probability sampling methods include accidental sampling, quota sampling and purposive sampling. In addition, non-response effects may turn *any* probability design into a non-probability design if the characteristics of non-response are not well understood, since non response effectively modifies each element's probability of being sampled.

12.3. SAMPLING TECHNIQUES

Within any of the types of frame identified above, a variety of sampling methods can be employed, individually or in combination. Factors commonly influencing the choice between these designs include:

- Nature and quality of the frame
- Availability of auxiliary information about units on the frame
- Accuracy requirements, and the need to measure accuracy
- Whether detailed analysis of the sample is expected
- Cost/operational concerns

12.3.1. PROBABILITY OR RANDOM SAMPLING

Probability sampling is based on the theory of probability. It is also known as random sampling. It provides a known nonzero chance of selection for each population element. It is used when generalization is the objective of study, and a greater degree of accuracy of estimation of population parameters is required. The cost and time required is high hence the benefit derived from it should justify the costs.

12.3.2 SIMPLE RANDOM SAMPLING

This sampling technique gives each element an equal and independent chance of being selected. An equal chance means equal probability of selection. An independent chance means that the draw of one element will not affect the chances of other elements being selected. The procedure of drawing a simple random sample consists of enumeration of all elements in the population.

1. Preparation of a List of all elements, giving them numbers in serial order 1, 2, B, and so on, and
2. Drawing sample numbers by using (a) lottery method, (b) a table of random numbers or (c) a computer.

Suitability: This type of sampling is suited for a small homogeneous population.

12.3.3 STRATIFIED SAMPLING

Where the population embraces a number of distinct categories, the frame can be organized by these categories into separate "strata." Each stratum is then sampled as an independent sub population, out of which individual elements can be randomly selected. There are several potential benefits to stratified sampling.

First, dividing the population into distinct, independent strata can enable researchers to draw inferences about specific subgroups that may be lost in a more generalized random sample.

Second, utilizing a stratified sampling method can lead to more efficient statistical estimates (provided that strata are selected based upon relevance to the criterion in question, instead of availability of the samples). Even if a stratified sampling approach does not lead to increased statistical efficiency, such a tactic will not result in less efficiency than would simple random sampling, provided that each stratum is proportional to the group's size in the population.

Third, it is sometimes the case that data are more readily available for individual, pre-existing strata within a population than for the overall population; in such cases, using a stratified sampling approach may be more convenient than aggregating data across groups (though this may potentially be at odds with the previously noted importance of utilizing criterion-relevant strata).

Finally, since each stratum is treated as an independent population, different sampling approaches can be applied to different strata, potentially enabling researchers to use the approach best suited (or most cost-effective) for each identified subgroup within the population.

There are, however, some potential drawbacks to using stratified sampling. First, identifying strata and implementing such an approach can increase the cost and complexity of sample selection, as well as leading to increased complexity of population estimates. Second, when examining multiple criteria, stratifying variables may be related to some, but not to others, further complicating the design, and potentially reducing the utility of the strata. Finally, in some cases (such as designs with a large number of strata, or those with a specified minimum sample size per group), stratified sampling can potentially require a larger sample than would other methods (although in most cases, the required sample size would be no larger than would be required for simple random sampling.

Advantages: Stratified random sampling enhances the representativeness to each sample, gives higher statistical efficiency, easy to carry out, and gives a self-weighing sample.

Disadvantages: A prior knowledge of the composition of the population and the distribution of the population, it is very expensive in time and money and identification of the strata may lead to classification of errors.

12.3.4 SYSTEMATIC SAMPLING

Systematic sampling relies on arranging the study population according to some ordering scheme and then selecting elements at regular intervals through that ordered list. Systematic sampling involves a random start and then proceeds with the selection of every k^{th} element from then onwards. In this case, k = (population size/sample size). It is important that the starting point is not automatically the first in the list, but is instead

randomly chosen from within the first to the k^{th} element in the list. A simple example would be to select every 10^{th} name from the telephone directory (an 'every 10^{th}' sample, also referred to as 'sampling with a skip of 10').

Suitability: Systematic selection can be applied to various populations such as students in a class, houses in a street, telephone directory etc.

Advantages: The advantages are it is simpler than random sampling, easy to use, easy to instruct, requires less time, it's cheaper, easier to check, sample is spread evenly over the population, and it is statistically more efficient.

Disadvantages: The disadvantages are it ignores all elements between two k^{th} elements selected, each element does not have equal chance of being selected, and this method sometimes gives a biased sample.

12.3.5 CLUSTER SAMPLING

Sometimes it is more cost-effective to select respondents in groups ('clusters'). Sampling is often clustered by geography, or by time periods. (Nearly all samples are in some sense 'clustered' in time - although this is rarely taken into account in the analysis.) For instance, if surveying households within a city, we might choose to select 100 city blocks and then interview every household within the selected blocks.

Clustering can reduce travel and administrative costs. In the example above, an interviewer can make a single trip to visit several households in one block, rather than having to drive to a different block for each household.

Suitability: The application of cluster sampling is extensive in farm management surveys, socio-economic surveys, rural credit surveys, demographic studies, ecological studies, public opinion polls, and large scale surveys of political and social behaviour, attitude surveys and so on.

Advantages: The advantages of this method is it is easier and more convenient, cost of this is much less, promotes the convenience of field work as it could be done in compact places, it does not require more time, units of study can be readily substituted for other units and it is more flexible.

Disadvantages: The cluster sizes may vary and this variation could increase the bias of the resulting sample. The sampling error in this method of sampling is greater and the adjacent units of study tend to have more similar characteristics than do units distantly apart.

12.3.6 AREA SAMPLING

This is an important form of cluster sampling. In larger field surveys cluster consisting of specific geographical areas like districts, talluks, villages or blocks in a city are randomly drawn. As the geographical areas are selected as sampling units in such cases, their sampling is called area sampling. It is not a separate method of sampling, but forms part of cluster sampling.

12.3.7 PROBABILITY-PROPORTIONAL-TO-SIZE SAMPLING

In some cases the sample designer has access to an "auxiliary variable" or "size measure", believed to be correlated to the variable of interest, for each element in the population. These data can be used to improve accuracy in sample design. One option is to use the auxiliary variable as a basis for stratification, as discussed above.

Another option is probability-proportional-to-size ('PPS') sampling, in which the selection probability for each element is set to be proportional to its size measure, up to a maximum of 1.

In a simple PPS design, these selection probabilities can then be used as the basis for Poisson sampling. However, this has the drawback of variable sample size, and different portions of the population may still be over- or under-represented due to chance variation in selections. To address this problem, PPS may be combined with a systematic approach.

Example: Suppose we have six schools with populations of 150, 180, 200, 220, 260, and 490students respectively (total 1500 students), and we want to use student population as the basis fora PPS sample of size three. To do this, we could allocate the first school numbers 1 to 150, the second school 151 to 330 (= 150 + 180), the third school 331 to 530, and so on to the last school(1011 to 1500). We then generate a random start between 1 and 500 (equal to 1500/3) and count through the school populations by multiples of 500. If our random start was 137, we would select the schools which have been allocated numbers 137, 637, and 1137, i.e. the first, fourth, and sixth schools.

Advantages: The advantages are clusters of various sizes get proportionate representation, PPS leads to greater precision than would a simple random sample of clusters and a constant sampling fraction at the second stage, equal-sized samples from each selected primary cluster are convenient for field work.

Disadvantages: PPS cannot be used if the sizes of the primary sampling clusters are not known.

12.3.8 DOUBLE SAMPLING AND MULTIPHASE SAMPLING

Double sampling refers to the subsection of the final sample form a preselected larger sample that provided information for improving the final selection. When the procedure is extended to more than two phases of selection, it is then, called multi-phase sampling. This is also known as sequential sampling, as sub-sampling is done from a main sample in phases. Double sampling or multiphase sampling is a compromise solution for a dilemma posed by undesirable extremes. "The statistics based on the sample of 'n' can be improved by using ancillary information from a wide base: but this is too costly to obtain from the entire population of N elements. Instead, information is obtained from a larger preliminary sample n_L which includes the final sample n.

12.3.9 NON-PROBABILITY OR NON RANDOM SAMPLING

Non-probability sampling or non-random sampling is not based on the theory of probability. This sampling does not provide a chance of selection to each population element.

Advantages: The only merits of this type of sampling are simplicity, convenience and low cost.

Disadvantages: The demerits are it does not ensure a selection chance to each population unit. The selection probability sample may not be a representative one. The selection probability is unknown. It suffers from sampling bias which will distort results.

12.3.10 QUOTA SAMPLING

In **quota sampling**, the population is first segmented into mutually exclusive sub-groups, just a sin stratified sampling. Then judgement is used to select the subjects or units from each segment based on a specified proportion. For example, an interviewer may be told to sample 200 females and 300 males between the age of 45 and 60.

It is this second step which makes the technique one of non-probability sampling. In quota sampling the selection of the sample is non-random. For example interviewers might be tempted to interview those who look most helpful. The problem is that these samples may be biased because not everyone gets a chance of selection. This random element is its greatest weakness and quota versus probability has been a matter of controversy for several years.

Suitability: It is used in studies like marketing surveys, opinion polls, and readership surveys which do not aim at precision, but to get quickly some crude results.

Advantage: It is less costly, takes less time, non-need for a list of population, and field work can easily be organized.

Disadvantage: It is impossible to estimate sampling error, strict control if field work is difficult, and subject to a higher degree of classification.

12.3.11 CONVENIENCE OR ACCIDENTAL SAMPLING

Accidental sampling (sometimes known as **grab**, **convenience** or **opportunity sampling**) is a type of non-probability sampling which involves the sample being drawn from that part of the population which is close to hand. That is, a population is selected because it is readily available and convenient. It may be through meeting the person or including a person in the sample when one meets them or chosen by finding them through technological means such as the internet or through phone. The researcher using such a sample cannot scientifically make generalizations about the total population from this sample because it would not be representative enough. For example, if the interviewer were to conduct such a survey at a shopping center early in the morning on a given day, the people that he/she could interview would be limited to those given there at that given time, which would not represent the views of other members of society in such an area, if the survey were to be conducted at different times of day and several times per week.

Suitability: Though this type of sampling has no status, it may be used for simple purposes such as testing ideas or gaining ideas or rough impression about a subject of interest.

Advantage: It is the cheapest and simplest, it does not require a list of population and it does not require any statistical expertise.

Disadvantage: The disadvantage is that it is highly biased because of researcher's subjectivity, it is the least reliable sampling method and the findings cannot be generalized.

12.3.12 PURPOSIVE (OR JUDGMENT) SAMPLING

This method means deliberate selection of sample units that conform to some pre-determined criteria. This is also known as judgment sampling. This involves selection of cases which we judge as the most appropriate ones for the given study. It is based on the judgment of the researcher or some expert. It does not aim at securing a cross section of a population. The chance that a particular case be selected for the sample depends on the subjective judgment of the researcher.

Suitability: This is used when what is important is the typicality and specific relevance of the sampling units to the study and not their overall representativeness to the population.

Advantage: It is less costly and more convenient and guarantees inclusion of relevant elements in the sample.

Disadvantage: It is less efficient for generalizing, does not ensure the representativeness, requires more prior extensive

12.3.13 SNOW-BALL SAMPLING

This is the colourful name for a technique of Building up a list or a sample of a special population by using an initial set of its members as informants. This sampling technique may also be used in socio-metric studies.

Suitability: It is very useful in studying social groups, informal groups in a formal organization, and diffusion of information among professional of various kinds.

Advantage: It is useful for smaller populations for which no frames are readily available.

Disadvantage: The disadvantage is that it does not allow the use of probability statistical methods. It is difficult to apply when the population is large. It does not ensure the inclusion of all the elements in the list.

12.4 SUMMARY

Sampling techniques help us in this situation. Sampling method is the only method that can be used in certain cases. There are some cases in which the census method is inapplicable and the only practicable means is provided by the sample method. Despite various advantages of sampling, it is not completely free from limitations. A sample survey must be carefully planned and executed otherwise the results obtained may be inaccurate and misleading. Even if a complete count care is taken still serious errors may arise in sampling, if the sampling procedure is not perfect. Sampling generally requires the services of experts, even only for

consultation purposes. In the absence of qualified and experienced persons, the information obtained from sample surveys cannot be relied upon. Shortage of experts in the sampling field is a serious hurdle in the way of reliable statistics. Sampling techniques may be classified into two broad categories namely probability and non-probability sampling. Non-probability sampling methods are those which do not provide every item in the universe with a known chance of being included in the sample. These include judgment, quota, cluster and convenience sampling techniques.

12.5 SELF-ASSESSMENT QUESTIONS

1. Define Probability and non-probability sampling
2. Elaborate and discuss the sampling techniques.
3. What are the advantages and disadvantages of stratified sampling?
4. Define:

a. Cluster sampling

b. Quota sampling

c. Accidental sampling

5. Critically examine the various probability sampling methods.
6. Distinguish between random sampling, purposive sampling and stratified sampling. How is a random sample obtained?

12.6 TEXT AND REFERENCES

- Donald Cooper, Donald R. Cooper, and Pamela S. Schindler, "Business Research Methods", Tata McGraw Hill Inc.
- Alan Bryman, and Emma Bell, "Business research Methods". Oxford University Press Publication.
- Donald S. Tull and Del S. Hawkins, "Marketing Research – Measurement and Method", Prentice Hall of India Pvt. Ltd., New Delhi.
- William G. Zikmund, "Business Research Methods", Thomson, South-Western Publication, Singapore.
- C.R. Kothari, "Research Methodology – Methods and Techniques", Wilely Eastern Limited, Delhi.
- S.P. Gupta, "*Statistical methods*" Sultan Chand & Sons publication, New Delhi
- Saunders, "Research Methods for Business Students", Pearsons Education Publications.
- Naresh Malhotra, "Marketing Research: An applied Orientation", Prentice Hall International Edition.
- T.S. Wilkinson, and P.L. Bhandarker, "Methodology and techniques of Social Research", Himalaya Publishing House, New Delhi.

UNIT 13: SOURCES OF DATA COLLECTION

UNIT FRAMEWORK

13.1 OBJECTIVE

After reading this unit you will be able to:

- Understand the meaning of data collection.
- Explain the importance of data
- Identify the various types of data.
- Explain from where the data is collected.
- Understand the importance of primary and secondary data
- Explain advantages and disadvantages of primary and secondary data.

13.2 INTRODUCTION: MEANING AND IMPORTANCE OF DATA

DEFINITION

"Data are facts, figures and other relevant materials past and present serving as bases for study and analysis".

MEANING OF DATA

The search for answers to research questions calls collection of Data. "Data are facts, figures and other

relevant materials, past and present, serving as bases for study and analysis".

TYPES OF DATA

The Data needed for social science Research may be broadly classified into:

A. Data pertaining to human beings

B. Data relating to organizations

C. Data pertaining to territorial area.

A) Personal Data (relating to Human beings) are of two types.

(a). Demographic and socio-economic characteristics of individuals. Like name, sex, race, social class, relation, education, occupation, income etc.

(b). Behavioural Variables: Attitudes, opinion knowledge, practice, intensions etc.

B) Organisation Data: - Consist of data relating to an organizations, origin ownership, function, performance etc.

C) Territorial Data: - are related to geo-physical characteristic, population, infrastructure etc. of divisions like villages, cities, taluks, distinct, state etc.

IMPORTANCE OF DATA

The data serve as the bases or raw materials for analysis without Data no specific inferences can be drawn on our study. Inferences based on imagination or guesswork cannot provide correct answers to research questions. The relevance, adequacy and reliability of data determine the quality of the findings of a study. The reliability of data determines the quality of research.

Data form the basis for testing the hypotheses formulated in a study. Data also provide the facts and figures required for constructing measurement scales and tables, which are analysed with statistical techniques. Inferences on the results of statistical, analysis and tests of significance provide the answers to research questions. Thus the scientific process of measurement, analysis, testing and inferences depends on the availability of relevant data and their accuracy. Hence the importance of data for any research studies.

SOURCES OF DATA

The sources of data may be classified into a) primary sources b) secondary sources. Both the sources of information have their merits and demerits. The selection of a particular source depends upon the (a) purpose and scope of enquiry, (b) availability of time, (c) availability of finance, (d) accuracy required, (e) statistical tools to be used, (f) sources of information (data), and (g) method of data collection.

13.3 PRIMARY SOURCES OF DATA

Primary sources are original sources from which the researcher directly collects data that have not been previously collected e.g., collection of data directly by the researcher on brand awareness, brand preference, brand loyalty and other aspects of consumer behaviour from a sample of consumers by interviewing them. Primary data are firsthand information collected through various methods such as observation, interviewing, mailing etc.

According to **P. V. Young**, "primary sources are those data gathered at first hand and the responsibility so of their compilation and promulgations remaining under the same authority that originally gathered them."

In the words of **Watter R. Borg**, "Primary sources are direct describing occurrences by an individual who actually observed on witness for occurrences."

13.3.1 ADVANTAGE OF PRIMARY DATA

- It is original source of data
- It is possible to capture the changes occurring in the course of time.
- It flexible to the advantage of researcher.
- Researchers know its accuracy.
- Only that data are collected which meet outs the objective of research project.
- In maximum methods of primary data collection researchers know who the respondents are so face to face communication is there.
- It is most authentic since the information is not filtered or tampered.
- Extensive research study is based of primary data

13.3.2 DISADVANTAGE OF PRIMARY DATA

- Primary data is expensive to obtain
- It is time consuming
- It requires extensive research personnel who are skilled.
- It is difficult to administer.
- Chances of biasness are at great extent.
- Biasness can also be there on the part of respondent. Wrong answer can be given by then which may affect the accuracy of data.
- It may have narrow coverage. It means researchers may collect data only within his/her reach or according to his mindset.

13.3.3 METHODS OF COLLECTING PRIMARY DATA

The following are the methods of collecting Primary Data:

1. Personal Interview
2. Observation Method
3. Data Collected Through Mail
4. Web Method
5. Telephone Method

1. PERSONAL INTERVIEW: Under this method the researcher personally visits the area of enquiry, establishes personal contact with the respondent and collects necessary facts and figures.

Example: If a researcher wants to know about the family income of persons in a particular area, he goes personally to the area and collects data on the basis of personal contacts. It will be direct personal interview.

Advantages:

The main advantages of personal interview are:

(i) **Generally yields highest cooperation and lowest refusal rates:** Since personal visits are made to

the respondents, to the refusal rates are low as better explanation about the research work can be given to the respondent.

(ii) **Allows for longer, more complex interviews:** Since contact is made personally and the researcher has time to explain the about the complexity of the research, in details, to the respondent, so he may go for complex interviews where questions have to be framed at the point of the respondent.

(iii) **High response quality:** The response quality of the respondent can be judged at the time of collecting the data, so the quality of data is controlled.

(iv) **Multi-method data collection:** Under this method the person collecting the data may change the method of data collection if he is not getting appropriate data by from his existing method.

Disadvantages

The main disadvantages of personal interview are:

(i) Most costly mode of administration: Since personal visits are made for conducting the interview and also two respondents may be at far away distance, so the cost of collecting the data increases, as it will include the transportation cost of the person collecting the data.

(ii) Longer data collection period: Since the data is collected by personal visit, so it becomes a very slow process to collect data. It" takes comparative longer period to collect the data under this method.

(iii) Interviewer concerns: Since the interview is conducted by making personal visits, so there is always a chance, that the respondent may not respond to the process, or may give his own suggestion for changes in the research and provide data according to the changes.

2. OBSERVATION METHOD: Under this method, the researcher collects information directly through systematic watching and noting the phenomena as they occur in nature with regard to cause and effect or mutual relations rather than through the reports of others. It is a process of recording relevant information without asking anyone specific questions and in some cases, even without the knowledge of the respondents.

Advantages:

(i.) **The respondent will provide data:** In any case the data is collected from the respondent by observing the response pattern and the respondents are unable or reluctant to provide information.

(ii.) **Data Accuracy:** The method provides deeper insights into the problem and generally the data is accurate and quicker to process. Therefore, this is useful for intensive study rather than extensive study.

Disadvantages

(i.) Unable to predict the occurrence of data: In many situations, the researcher cannot predict when the events will occur. So when an event occurs there may not be a ready observer to observe the event.

(ii.) No true data: As the respondent may be aware of the observer and as a result may alter their behavioural pattern.

(iii.) Paradigm: Since this method solely depends on the observation power of the researcher, so due to lack of training and paradi1 the researcher may not observe the things as they occur.

(iv.) Not suitable for large research: This method cannot be used extensively if the inquiry is large and spread over a wide area.

3. DATA COLLECTED THROUGH MAIL: Under this method the data a collected by sending letters to the respondent. A letter may contain questionnaire and the respondent is required to respond back.

Advantages

(i.) **Generally lowest cost:** As compared to other form of data collection, this method is cheapest as the questionnaire maximum of one page. Further to cut the cost, the researcher may go through print media to get its questionnaire distributed to the respondent.

(ii.) **Can be administered by smaller team of people:** This research can be administered with few staffs as only the office staffs are required and no field staffs are necessary.

(iii.) **Access to otherwise difficult to locate, busy populations:** As the research is done through mail, so the researcher could get the data from the respondent otherwise it is difficult to locate amongst the busy population.

(iv.) **Respondents can look up information or consult with others:** As the respondent don't have to respond to the queries of the researcher, so the respondent gets enough time to understand the information required and he may even consult other person to provide the accurate data.

Disadvantages

(i.) **Most difficult to obtain cooperation:** As the respondent has the option to respond back or not, so the researcher may find it difficult to collect information as all of its questionnaire sent may not return with the data. Also there is problem of delivery of mails to the respondent.

(ii.) **No researcher involved in collection of data:** As all the work of collection of data are through mails, so there are no researcher involved in the process, so the respondent may sometimes find it difficult to understand the queries raised by the researcher.

(iii.) **Need good sample:** As the mails are sent with the help of some database, so the researcher may not know, how the respondent will react? The respondent mayor may not provide appropriate data as required by the researcher.

(iv.) **More likely to need an incentive for respondent:** In order to make respondent give response to the letter sent, there should be some incentive scheme to be attached with the mail, otherwise, the response rate will be very poor.

(v.) **Slower data collection period:** As the time needed for delivery of mails to the respondent and back, so this method is much slower than how the data is collected through telephone or personal interview or web mail.

4. WEB METHOD: Under this method, the data is collected through internet. This method can be further divided into two groups

A. **Through Polling:** The researcher may put the information on a web server and the respondent may require responding to the information through online poll, or blog.

B. **Through Mails:** The researcher may also opt to send emails to various respondents and may give them the option to respond back.

Advantages

(i.) **Lower cost:** As the research work is completed online and the data are collected in the database, so the researcher may not require any paper, postage, mailing, data entry costs.

(ii.) **Can reach international populations:** As the research work is done online, so the researcher may also involve international respondent for collection of data at no extra cost or effort.

(iii.) **Time required for implementation reduced:** As the respondent is required to send the response online, and the data is also collected, so the researcher don't have to waste time in compiling the data and interpreting the data. The researcher may directly go for interpretation.

(iv.) **Complex patterns can be programmed:** As the research work is online, so the researcher may go for complex research activity as all the queries of the respondent are immediately handled by the researcher.

(v.) **Sample size can be greater:** As the research work is online, the sample size for the research may be greater as the population outside the country can be included.

Disadvantages

(i.) **Limitation of technology:** As in India, approximately 55% of homes own a computer; 30% have home e-mail, so the choice of population is restricted for the researcher.

(ii.) **Representative samples difficult:** Since the access to the technology is difficult for general population, so the data collection activity cannot generate random samples of the population.

(iii.) **Differences in capabilities of people's computers and software for accessing Web surveys:** Since each person has different capabilities and knowledge about the usage and utility of the web, so a good respondent may not provide the sample for the research.

(iv.) **Difference in people's response:** The researcher cannot ascertain, if the same person has given response to the survey. If a research activity is performed by some other person, then the quality of response will be different.

(v.) **Different ISPs/line speeds limits extent of graphics that can be used:** If the researcher is using graphic display to explain the theme or complexities of his research to the respondent, so it is quite possible that some of the respondent may not get the graphics displayed on their computer due to ISPs/line speeds limit or restriction of usage. In' such case the data collected will not be accurate.

5. TELEPHONE METHOD: Under this method, the researcher calls the respondent and collects the data over the telephone. He may use the telephone numbers available on the Telephone directory, and select the samples from the given population.

Advantages

(i.) **Less expensive than personal interviews:** As the research can be completed over the phone, so this is less expensive than the personal interview and that the ·data are collected quickly.

(ii.) **Samples from general population:** As telephones are accessible to the general population, than the Web Method, so more samples can be collected from larger population.

(iii.) **Shorter data collection period than personal interviews:** As limited tools can be used to explain the research objective to the respondent, so it takes much lesser time to collect data. Also the larger population can be reached in very short time.

(iv.) **Researcher administration:** As the researcher can explain and listen to the queries of the respondent, so in this method there is direct control of the researcher over the subject for data collection than the mails or web mails.

(v.) **Better control and supervision of Researcher:** Similarly, as the researcher is contacting the respondent from his place, so the researcher may refer any literature during the process of collecting data. This is restricted in case of personal interview.

(vi.) **Better response rate than mail for list samples:** For Comparative Scaling techniques this method is easy and effective as the researcher will provide the list and respondent have to choose from the available lists.

Disadvantages:

(i.) **Biased against households without telephones, unlisted numbers:** As still most homes in India,

don't have telephone numbers, or their numbers are not listed in the telephone directories due to having pre-paid connections. In such case such persons don't get equal opportunities to appear for research activity.

(ii.) **No response:** If the respondents don't pick up the telephone on time or if the call is missed, in such case the researcher may select any other person as sample.

(iii.) **Questionnaire constraints:** Complex questionnaire prepared by the researcher cannot be used in this method, if next question of the questionnaire depends on the answer of the respondent then such types of questionnaire cannot be included in the questionnaire.

(iv.) **Difficult to administer questionnaires on sensitive or complex topics:** Since the researcher is getting the names of the samples from the telephone directory, and he is not aware about the economic, social or emotional situation of the samples, so it is very difficult to administer questionnaires on sensitive or complex topics.

13.4 SECONDARY SOURCES OF DATA

These are sources containing data which have been collected and compiled for another purpose. The secondary sources consists of readily compendia and already compiled statistical statements and reports whose data may be used by researchers for their studies e.g., census reports , annual reports and financial statements of companies, Statistical statement, Reports of Government Departments, Annual reports of currency and finance published by the Reserve Bank of India, Statistical statements relating to Cooperatives and Regional Banks, published by the NABARD, Reports of the National sample survey Organization, Reports of trade associations, publications of international organizations such as UNO, IMF, World Bank, ILO, WHO, etc., Trade and Financial journals newspapers etc.

Secondary sources consist of not only published records and reports, but also unpublished records. The latter category includes various records and registers maintained by the firms and organizations, e.g., accounting and financial records, personnel records, register of members, minutes of meetings, inventory records etc.

13.4.1 FEATURES OF SECONDARY DATA

Though secondary sources are diverse and consist of all sorts of materials, they have certain common characteristics.

First, they are readymade and readily available, and do not require the trouble of constructing tools and administering them.

Second, they consist of data which a researcher has no original control over collection and classification. Both the form and the content of secondary sources are shaped by others. Clearly, this is a feature which can limit the research value of secondary sources.

Finally, secondary sources are not limited in time and space. That is, the researcher using them need not have been present when and where they were gathered.

13.4.2 USES OF SECONDARY DATA

The second data may be used in three ways by a researcher. First, some specific information from secondary sources may be used for reference purpose. For example, the general statistical information in the number of co-operative credit societies in the country, their coverage of villages, their capital structure, volume of business etc., may be taken from published reports and quoted as background information in a study on the evaluation of performance of cooperative credit societies in a selected district/state.

Second, secondary data may be used as bench marks against which the findings of research may be tested, e.g., the findings of a local or regional survey may be compared with the national averages; the performance indicators of a particular bank may be tested against the corresponding indicators of the banking industry as a whole; and so on.

Finally, secondary data may be used as the sole source of information for a research project. Such studies as securities Market Behaviour, Financial Analysis of companies, Trade in credit allocation in commercial banks, sociological studies on crimes, historical studies, and the like, depend primarily on secondary data. Year books, statistical reports of government departments, report of public organizations of Bureau of Public Enterprises, Censes Reports etc, serve as major data sources for such research studies.

13.4.3 ADVANTAGES OF SECONDARY DATA

Secondary sources have some advantages:

1. Secondary data, if available can be secured quickly and cheaply. Once their source of documents and reports are located, collection of data is just matter of desk work. Even the tediousness of copying the data from the source can now be avoided, thanks to Xeroxing facilities.

2. Wider geographical area and longer reference period may be covered without much cost. Thus, the use of secondary data extends the researcher's space and time reach.

3. The use of secondary data broadens the data base from which scientific generalizations can be made.

4. Environmental and cultural settings are required for the study.

5. The use of secondary data enables a researcher to verify the findings bases on primary data. It readily meets the need for additional empirical support. The researcher need not wait the time when additional primary data can be collected.

13.4.4 DISADVANTAGES OF SECONDARY DATA

Although secondary data are easy to access and cost-effective, they also have significant limitations:

1. The secondary data are not up-to-date and become obsolete when they appear in print, because of time lag in producing them. For example, population census data are published two or three years later after compilation and no new figures will be available for another ten years.

2. Data may be too broad-based that is, not specific enough to adequately address the firm's research questions.

3. The units in which the data are presented may not be meaningful.

4. The source of the data may not provide sufficient supporting material to allow the researcher to judge the quality of the research.

5. The data sources may lack reliability and credibility. Some secondary data may simply be inaccurate.

6. The most important limitation is the available data may not meet our specific needs. The definitions adopted by those who collected those data may be different; units of measure may not match; and time periods may also be different.

7. The available data may not be as accurate as desired. To assess their accuracy we need to know how the data were collected.

8. Finally, information about the whereabouts of sources may not be available to all social scientists. Even if the location of the source is known, the accessibility depends primarily on proximity. For example, most of the unpublished official records and compilations are located in the capital city, and they are not within the easy reach of researchers based in far off places.

13.4.5 METHODS OF COLLECTING SECONDARY DATA

The researcher may get Secondary data from two sources (a) Published; (b) Unpublished. While we are going to discuss about published data, the unpublished data may be in records of Government or private organizations, research organizations, research scholars etc. However, these data being

unpublished is not freely available and the details about the sources remain with few persons.

Most of the researcher use published data as they are available from the following sources:

1. **Newspaper and Magazines:** Statistical data on a number of current socioeconomic subjects can be obtained from data collected and published by some reputed newspapers, magazines, periodicals, etc.

2. **Published Articles of an individual:** Many times some research studies are carried on at individual level and are published in magazines in the form of articles or in the form of books. Though these studies are based on research works of limited area, however they may provide useful information for further research.

3. **Government Publication and Gazetteer:** Various ministries and departments of Central Government and State Governments publish data regularly on a number of subjects. These data are considered reasonably reliable for research work.

4. **Reports from Commissions and Committees:** The government constitutes committees and commissions for the study and enquiry of various problems from time to time, which submit their reports after collection and analysis of required data and information. The information contained in such reports is very useful and reliable.

5. **Publications from various Organizations:** Many important universities and semi-government research organizations also publish their research studies and these publications are important sources for analytical statistical information.

6. **Private Data Publication:** Recent trends have emerged where the private organizations are collecting and compiling various data for further usage as research data.

13.5 DIFFERENCE BETWEEN PRIMARY DATA AND SECONDARY DATA

(a). Primary data are collected by the researcher himself, although the secondary data has been collected previously by other researcher.

(b). Primary data are collected and used first time. However, on the secondary data, some decisions has been made previously, such decisions mayor may not be useful for the researcher now.

(c). Since primary data is collected by the researcher himself, so it relates directly to the research objective *or* may be more close to research objective. How many parts of secondary data need not to be related to the research objective?

(d). Since primary data is collected by the researcher, so it is more time taking activity for the researcher than to get the secondary data.

The students should note here that the primary data and secondary data differ in the degree of impact only and that same set of data may be secondary in the hands of one and primary in the hands of another. As Secrist has said "the distinction between primary am secondary data is largely one of degree. Data which are secondary in the hands of one party may be primary in the hands of another" Example, if population statistics are primary which it is collected under population census but it becomes secondary when it is used for some researcher.

13.6 QUESTIONNAIRE AND SCHEDULE CONSTRUCTION

QUESTIONNAIRE

A **questionnaire** is a research instrument consisting of a series of questions and other prompts for the purpose of gathering information from respondents. Although they are often designed for statistical analysis of the responses, this is not always the case. The questionnaire was invented by Sir Francis Galton. Questionnaires have advantages over some other types of surveys in that they are cheap, do not require as much effort from

the questioner as verbal or telephone surveys, and often have standardized answers that make it simple to compile data.

However, such standardized answers may frustrate users. Questionnaires are also sharply limited by the fact that respondents must be able to read the questions and respond to them. Thus, for some demographic groups conducting a survey by questionnaire may not be practical. As a type of survey, questionnaires also have many of the same problems relating to question construction and wording that exist in other types of opinion polls.

Question Sequence: In general, questions should flow logically from one to the next. To achieve the best response rates, questions should flow from the least sensitive to the most sensitive, from the factual and behavioural to the attitudinal, and from the more general to the more specific. There typically is a flow that should be followed when constructing a questionnaire in regards to the order that the questions are asked. The order is as follows:

1. Screens

2. Warm-ups

3. Transitions

4. Skips

5. Difficult

6. Classification

Screens are used as a screening method to find out early whether or not someone should complete the questionnaire. **Warm-ups** are simple to answer, help capture interest in the survey, and may not even pertain to research objectives. **Transition** questions are used to make different areas flow well together. **Skips** include questions similar to "If yes, then answer question 3. If no, then continue to question 5." **Difficult** questions are towards the end because the respondent is in "response mode." Also, when completing an online questionnaire, the progress bars let the respondent know that they are almost done so they are more willing to answers more difficult questions. **Classification**, or demographic, question should be at the end because typically they can feel like personal questions which will make respondents uncomfortable and not willing to finish survey.

SCHEDULE

This method of data collection is very much like the collection of data through questionnaire with little difference which likes in the fact that schedules are being filled in by enumerators who are specially appointed for this purpose. These enumerators along with schedules go to respondents, put to them the questions from the performer in the order questions are listed and record the replay in the space meant for the same Performa. This method requires the selection and training of enumerators to fill up the schedules and they should be carefully selected. Enumerators should be intelligent and must be able to find out the truth. The enumerators should be honest sincere and hard working. This method is very useful because it yield good results. Population censes all over the world is conducted through this method.

DIFFERENCES BETWEEN SCHEDULE AND QUESTIONNAIRE

1. The Questionnaire is generally sent through mail to informants. The schedule is generally filled by the research worker.

2. To collect data through questionnaire is relatively cheap. To collect data through schedule is relatively more expensive.

3. Non- response is high in case of questionnaire whereas in schedule response is very high.

4. In Questionnaire there is no personal conducts. But in a schedule there is a face-to face contact.

5. The questionnaire method is used only when respondents are literate.
6. Along with schedules observation methods can be also used.

13.7 BASIC RULES FOR QUESTIONNAIRE ITEM CONSTRUCTION

- Use statements which are interpreted in the same way by members of different subpopulations of the population of interest.
- Use statements where persons that have different opinions or traits will give different answers.
- Think of having an "open" answer category after a list of possible answers.
- Use only one aspect of the construct you are interested in per item.
- Use positive statements and avoid negatives or double negatives.
- Do not make assumptions about the respondent.
- Use clear and comprehensible wording, easily understandable for all educational levels
- Use correct spelling, grammar and punctuation.
- Avoid items that contain more than one question per item (e.g. Do you like strawberries and potatoes?).

Concerns with questionnaires

While questionnaires are inexpensive, quick, and easy to analyze, often the questionnaire can have more problems than benefits. For example, unlike interviews, the people conducting the research may never know if the respondent understood the question that was being asked. Also, because the questions are so specific to what the researchers are asking, the information gained can be minimal.

Often, questionnaires such as the Myers-Briggs Type Indicator, give too few options to answer; respondents can answer either option but must choose only one response. Questionnaires also produce very low return rates, whether they are mail or online questionnaires. The other problem associated with return rates is that often the people that do return the questionnaire are those that have a really positive or a really negative viewpoint and want their opinion heard. The people that are most likely unbiased either way typically don't respond because it is not worth their time.

13.8 SUMMARY

Data are facts and other relevant materials, past and present, serving as bases for study and analyses. The data needed for a social science research may be broadly classified into (a) Data pertaining to human beings, (b) Data relating to organization and (c) Data pertaining to territorial areas. Personal data or data related to human beings consists of: Demographic and socio-economic characteristics of individuals: Age, sex, race, social class, religion, marital status, education, occupation income, family size, location of the household life style etc.

Data may broadly be divided into two categories, namely **primary data** and **secondary data.** The primary data are those which are collected for the first time by the organisation which is using them. The secondary data, on the other hand, are those which, have already been collected by some other agency but also can be used by the organisation under consideration. Primary data maybe collected by **observation, oral investigation, and questionnaire method** or by **telephone interviews.** Questionnaires may be used for data **collection by interviewers.** They may also be mailed to **prospective respondents.** The drafting of a good questionnaire requires utmost skill. The process of interviewing also requires a great deal of tact, patience and competence to establish rapport with the respondent. Secondary data are available in various published and unpublished documents. The suitability, reliability, adequacy and accuracy of the secondary data should,

however, be ensured before they are used for research problems.

It is always a tough task for the researcher to choose between primary and secondary data. Though primary data are more authentic and accurate, time, money and labor involved in obtaining these more often prompt the researcher to go for the secondary data. There are certain amount of doubt about its authenticity and suitability, but after the arrival of many government and semi government agencies and some private institutions in the field of data collection, most of the apprehensions in the mind of the researcher have been removed.

13.9 SELF-ASSESSMENT QUESTIONS

1. What are the types of data?
2. What are the primary sources of data?
3. Discuss the methods of collecting primary data.
4. How is personal interview done?
5. How is data collected through mail for doing research?
6. Write a short note on the following:

a. Telephone Method

b. Questionnaire

c. Schedule

7. How is questionnaire and schedule constructed?
8. What are the basic rules for questionnaire item construction?
9. What are the sources of secondary data?
10. How is secondary data useful to researcher?
11. What are the advantages of secondary data?
12. Describe the disadvantages of secondary data.
13. Discuss the methods of collecting secondary data.
14. Point out the difference between primary data and secondary data.

13.10 TEXT AND REFERENCES

- Hague Paul, “Market Research- a guide to planning, methodology & evaluation”; Kogan page, London.
- Khan, J.A.; “Research Methodology”; APH Publishing Corporation, New Delhi.
- Kothari, C.R.; “Research Methodology-Methods & Techniques”; New Age International (P) Limited.
- Kumar Rajendra ; “Research Methodology”; APH Publishing corporation, New Delhi.
- Kumar Ranjit; “Research methodology”; Pearson Education.
- Mehta, J.D. and Gupta, Umesh ; “Research Methods in Management”; Ramesh Book Depot, Jaipur-New Delhi.
- Mertens, Donna, M. ; “Research Methods in Education and Psychology”; Sage Publications, New Delhi.

- Murthy C.; “Research Methodology’; Vrinda Publications (P) Ltd. Delhi.
- Gopal, M.H. 1964. An Introduction to Research Procedure in Social Sciences, Asia Publishing House: Bombay.
- Sadhu, A.N. and A. Singh. 1980. Research Methodology in Social Sciences, Sterling Publishers Private Limited: New Delhi.
- Wilkinson, T.S. and P.L. Bliandarkar. 1979. Methodology and Techniques of Social Research, Himalaya Publishing House: Bombay.

BLOCK IV : STATISTICAL INVESTIGATION AND ESTIMATION

UNIT 14 STATISTICAL ESTIMATION

UNIT STRUCTURE

14.0 OBJECTIVES

After going through this unit you are able to know about

- The point and interval estimation
- Their uses advantages disadvantages and types.

14.1 INTRODUCTION

Statistical estimation is the process of using sample data to make informed guesses or estimates about a population parameter. It involves: Pint Estimation and Interval Estimations

14.2. POINT ESTIMATION:

Estimating a single value for a parameter. Point estimation is a statistical method that involves estimating a single value for a population parameter based on a sample of data. The goal is to provide a best guess or a single estimate for the parameter.

14.3 TYPES OF POINT ESTIMATORS:

In statistics, a point estimator is a function that takes a sample of data as input and returns a single value, known as a point estimate, that aims to estimate a population parameter. Here are some common types of point estimators:

1. Maximum Likelihood Estimator (MLE): Estimates the parameter value that maximizes the likelihood of observing the sample data.
2. Method of Moments Estimator (MME): Matches population moments (e.g., mean, variance) with sample moments.
3. Bayes Estimator: Uses Bayesian inference to estimate parameters, incorporating prior knowledge and sample data.
4. Least Squares Estimator (LSE): Minimizes the sum of squared errors between observed and predicted values.
5. Minimum Variance Unbiased Estimator (MVUE): Provides the most accurate estimate among all unbiased estimators.
6. Consistent Estimator: Converges to the true parameter value as sample size increases.
7. Unbiased Estimator: Averages out to the true parameter value over repeated sampling.
8. Biased Estimator: Systematically over- or underestimates the true parameter value.
9. Robust Estimator: Resistant to outliers and deviations from assumptions.
10. Non-Parametric Estimator: Doesn't assume a specific distribution for the data.

14.4 PROPERTIES OF POINT ESTIMATORS:

Point estimators have several important properties that help evaluate their quality and suitability. Here are the key properties of point estimators:

1. Unbiasedness: The expected value of the estimator equals the true population parameter.
2. Bias: The difference between the expected value of the estimator and the true population parameter.
3. Consistency: The estimator converges to the true population parameter as the sample size increases.
4. Efficiency: The estimator has the smallest variance among all unbiased estimators.

5. Sufficiency: The estimator uses all relevant information in the sample.
6. Robustness: The estimator is resistant to outliers, non-normality, and other deviations from assumptions.
7. Consistency: The estimator converges to the true population parameter as the sample size increases.
8. Asymptotic normality: The estimator's distribution approaches normality as the sample size increases.
9. Minimum variance: The estimator has the smallest variance among all estimators.
10. Invariance: The estimator's properties remain unchanged under transformations.
11. Completeness: The estimator uses all available information.
12. Ancillarity: The estimator's distribution doesn't depend on the parameter.

Desirable properties of point estimators include:

- Unbiasedness
- Consistency
- Efficiency
- Sufficiency
- Robustness

Undesirable properties include:

- Bias
- Inconsistency
- Inefficiency
- Lack of robustness

14.5 POINT ESTIMATION METHODS:

Here are common point estimation methods:

1. Maximum Likelihood Estimation (MLE): Estimates parameters that maximize the likelihood of observing the sample data.
2. Method of Moments (MOM): Matches population moments (mean, variance, etc.) with sample moments.
3. Least Squares Estimation (LSE): Minimizes the sum of squared errors between observed and predicted values.
4. Bayesian Estimation: Uses Bayesian inference to estimate parameters, incorporating prior knowledge and sample data.
5. Minimum Variance Unbiased Estimation (MVUE): Finds the unbiased estimator with the smallest variance.

7. Best Linear Unbiased Estimation (BLUE): Finds the linear unbiased estimator with the smallest variance.
8. Maximum a Posterior (MAP) Estimation: Estimates parameters that maximize the posterior distribution.
9. Minimum Mean Square Error (MMSE) Estimation: Minimizes the expected squared error between estimates and true values.
10. Pitman Estimation: Minimizes the expected squared error for a specific loss function.
11. Empirical Bayes Estimation: Combines Bayesian and frequents approaches.
12. Generalized Method of Moments (GMM): Extends MOM to handle non-linear models.
13. Simulated Minimum Distance (SMD) Estimation: Uses simulation to estimate parameters.
14. Indirect Inference Estimation: Estimates parameters by matching simulated and observed data.
15. Quasi-Maximum Likelihood Estimation (QMLE): Estimates parameters using a misspecified likelihood function.
16. Generalized Least Squares (GLS) Estimation: Accounts for heteroscedasticity and correlation.

14.6 ADVANTAGES OF POINT ESTIMATION:

Point estimation has several advantages:

1. Simplicity: Provides a single, easy-to-understand value as an estimate.
2. Easy interpretation: Directly estimates the population parameter.
3. Efficient: Often requires smaller sample sizes.
4. Computational ease: Many point estimators have closed-form solutions.
5. Widespread applicability: Suitable for various data types and distributions.
6. Confidence interval construction: Enables construction of confidence intervals.
7. Hypothesis testing: Facilitates hypothesis testing.
8. Model selection: Helps compare and select models.
9. Robustness: Some point estimators are robust to outliers and non-normality.
10. Theoretical foundation: Based on established statistical theories.
11. Easy communication: Simple to present and explain results.
12. Fast computation: Often faster than interval estimation or Bayesian methods.
13. Wide software availability: Most statistical software packages support point estimation.
14. Facilitates decision-making: Provides a single value for decision-making.
15. Building block for interval estimation: Point estimates are used to construct confidence intervals.

14.7 DISADVANTAGES OF POINT ESTIMATION:

Point estimation has several disadvantages:

1. Lack of precision: Provides a single value, masking variability.
2. Uncertainty: Doesn't quantify estimation uncertainty.
3. Risk of misinterpretation: May be mistaken for the true population parameter.
4. Sensitive to outliers: Some point estimators are affected by extreme values.
5. Assumption-dependent: Often relies on specific distributional assumptions.
6. Biased estimators: Some point estimators are biased, leading to inaccurate estimates.
7. Inefficient: May not use all available information.
8. Not suitable for all parameters: Some parameters are difficult to estimate using point estimation.
9. Ignores prior knowledge: Doesn't incorporate prior information (unlike Bayesian methods).
10. Difficult to quantify error: Doesn't provide a direct measure of estimation error.
11. Limited information: Provides only a single value, lacking additional insights.
12. Vulnerable to model misspecification: Assumes a correct model specification.
13. Can be misleading: May lead to incorrect conclusions if not interpreted carefully.
14. Not suitable for rare events: Point estimation may not capture rare events or outliers.
15. Requires large samples: Some point estimators require large sample sizes.

14.8 Errors in Point Estimation

Common point estimation errors:

1. Bias: Systematic deviation from the true value.
2. Variance: Excessive fluctuation in estimates.
3. Mean Squared Error (MSE): Combination of bias and variance.
4. Underestimation/Overestimation: Consistently estimating too low or too high.
5. Sampling Error: Errors due to sampling variability.
6. Non-Response Error: Errors due to missing data.
7. Measurement Error: Errors in data collection.
8. Model Misspecification: Incorrect model assumptions.
9. Confounding Variables: Ignoring influential variables.
10. Outliers: Extreme values affecting estimates.
11. Non-Random Sampling: Biased sampling methods.

12. Insufficient Sample Size: Too few data points.
13. Violating Assumptions: Ignoring statistical assumptions.
14. Numerical Errors: Calculation mistakes.
15. Interpretation Errors: Misunderstanding estimates.

To minimize point estimation errors:

1. Use appropriate estimation methods.
2. Check assumptions.
3. Evaluate estimator properties.
4. Consider multiple estimators.
5. Use robust methods.
6. Increase sample size.
7. Validate data.
8. Account for confounding variables.
9. Use model diagnostics.
10. Interpret estimates cautiously.

14.9 Uses of Point Estimation

Point estimation has numerous uses in various fields:

1. Population mean estimation: Estimate average values, like population means.
2. Proportion estimation: Estimate proportions, such as success rates.
3. Regression analysis: Estimate regression coefficients.
4. Hypothesis testing: Test hypotheses about population parameters.
5. Confidence interval construction: Create confidence intervals.
6. Prediction: Predict future values or outcomes.
7. Quality control: Monitor and control processes.
8. Reliability engineering: Estimate failure rates and reliability.
9. Survey research: Estimate population characteristics.
10. Medical research: Estimate treatment effects, disease prevalence.
11. Financial analysis: Estimate stock prices, portfolio risk.
12. Engineering: Estimate system parameters, signal processing.
13. Environmental studies: Estimate population sizes, growth rates.
14. Social sciences: Estimate population means, proportions.

15. Business: Estimate demand, sales, customer satisfaction.
16. Epidemiology: Estimate disease incidence, prevalence.
17. Pharmacokinetics: Estimate drug concentrations, absorption rates.
18. Image processing: Estimate image parameters.
19. Signal processing: Estimate signal parameters.
20. Machine learning: Estimate model parameters.

Point estimation is essential in:

- Data analysis
- Statistical inference
- Decision-making
- Modeling
- Forecasting
- Optimization

14.10. INTERVAL ESTIMATION:

Estimating a range of values for a parameter. Point estimation is a statistical method that involves estimating a single value for a population parameter based on a sample of data. The goal is to provide a best guess or a single estimate for the parameter.

14.11 TYPES OF INTERVAL ESTIMATION

Interval estimation provides a range of values within which a population parameter is likely to lie. There are several types of interval estimation:

1. Confidence Intervals: Estimate a population parameter with a specified level of confidence (e.g., 95%).
2. Prediction Intervals: Predict a future observation's value within a specified range.
3. Tolerance Intervals: Estimate a range of values containing a specified proportion of the population.
4. Bayesian Intervals (Credible Intervals): Estimate a range of values based on Bayesian inference.
5. Likelihood Intervals: Estimate a range of values based on likelihood theory.
6. Bootstrap Intervals: Estimate a range of values using re sampling techniques.
7. Jack knife Intervals: Estimate a range of values by systematically leaving out observations.
8. Profile Likelihood Intervals: Estimate a range of values using profile likelihood functions.
9. Fieller Intervals: Estimate a range of values for ratios of parameters.
10. Highest Density Intervals (HDIs): Estimate a range of values with the highest probability density.

11. Equi-Tailed Intervals: Estimate a range of values with equal tail probabilities.
12. Shortest Intervals: Estimate the shortest possible interval containing the parameter.

14.12 Properties of Interval estimation

Interval estimation has several important properties:

1. Coverage Probability: The probability that the interval contains the true parameter value.
2. Confidence Level (1 - α): The probability that the interval contains the true parameter value.
3. Width: The length of the interval, indicating precision.
4. Accuracy: Closeness of the interval to the true parameter value.
5. Precision: Inverse of the width, indicating how precise the interval is.
6. Reliability: Dependability of the interval in repeated sampling.
7. Un biasedness: Interval estimation is unbiased if the expected width is minimized.
8. Efficiency: Interval estimation is efficient if it has the smallest width among all unbiased intervals.
9. Consistency: Interval estimation is consistent if it converges to the true parameter value as sample size increases.
10. Robustness: Interval estimation is robust if it performs well under deviations from assumptions.
11. Invariance: Interval estimation is invariant if it remains unchanged under transformations.
12. Interpretability: Interval estimation provides a clear interpretation of the results.

14.13 INTERVAL ESTIMATION METHODS

Here are common interval estimation methods:

1. Confidence Intervals: Based on sample statistics and confidence levels.
2. Prediction Intervals: Estimate future observation ranges.
3. Tolerance Intervals: Contain a specified proportion of population values.
4. Bayesian Intervals (Credible Intervals): Use Bayesian inference and prior distributions.
5. Bootstrap Intervals: Re sampling-based intervals.
6. Jack knife Intervals: Systematically leave out observations.
7. Profile Likelihood Intervals: Use likelihood functions.
8. Fieller Intervals: Estimate ratios of parameters.
9. Highest Density Intervals (HDIs): Highest probability density ranges.
10. Equi-Tailed Intervals: Equal tail probabilities.
11. Shortest Intervals: Minimum width intervals.

12. Likelihood Ratio Intervals: Use likelihood ratio tests.
13. Wald Intervals: Use Wald tests and standard errors.
14. Score Intervals: Use score tests.
15. Pivotal Intervals: Use pivotal quantities.
16. Generalized Confidence Intervals: For complex models.
17. Simultaneous Intervals: Multiple parameters or comparisons.
18. Sequential Intervals: Monitor and update intervals.
19. Adaptive Intervals: Adjust interval width based on data.
20. Non-Parametric Intervals: Distribution-free methods.

14.14 ADVANTAGES OF INTERVAL ESTIMATION

Interval estimation offers several advantages:

1. Quantifies uncertainty: Provides a range of values, acknowledging estimation uncertainty.
2. More informative: Offers more information than point estimates alone.
3. Confidence level: Specifies the probability of containing the true parameter.
4. Flexibility: Various methods for different data types and assumptions.
5. Robustness: Some methods are robust to outliers and non-normality.
6. Hypothesis testing: Facilitates hypothesis testing and decision-making.
7. Model selection: Helps compare and select models.
8. Prediction: Enables prediction intervals for future observations.
9. Tolerance intervals: Estimates ranges for specific proportions.
10. Bayesian inference: Incorporates prior knowledge and uncertainty.
11. Improved interpretation: Enhances understanding of results.
12. Communication: Facilitates clear communication of findings.
13. Decision-making: Supports informed decision-making.
14. Uncertainty analysis: Enables uncertainty analysis and propagation.
15. Compliance: Meets regulatory requirements in some fields.
16. Sample size determination: Helps determine required sample sizes.
17. Power analysis: Facilitates power analysis and study design.
18. Meta-analysis: Enables combining results from multiple studies.
19. Subgroup analysis: Facilitates subgroup comparisons.
20. Graphical representation: Allows for intuitive graphical representation.

14.15 Disadvantages of Interval estimation

Interval estimation also has some disadvantages:

1. Width uncertainty: Interval width may be too wide or narrow.
2. Complexity: Some methods can be mathematically complex.
3. Assumption dependence: Many methods rely on assumptions (e.g., normality).
4. Interpretation challenges: Requires understanding of statistical concepts.
5. Overlapping intervals: Intervals may overlap, leading to ambiguous conclusions.
6. Confidence level selection: Choosing appropriate confidence levels can be subjective.
7. Sample size requirements: Large sample sizes may be needed for precise intervals.
8. Computational intensity: Some methods require extensive computations.
9. Software limitations: Not all software packages support advanced interval estimation methods.
10. Misinterpretation risks: Intervals may be misinterpreted as probability ranges.
11. Comparison difficulties: Comparing intervals across studies or groups can be challenging.
12. Model misspecification: Intervals may not account for model misspecification.
13. Non-coverage risks: Intervals may not contain the true parameter value.
14. Interval construction challenges: Constructing intervals for complex parameters can be difficult.
15. Prior distribution specification: Bayesian intervals require specifying prior distributions.
16. Computational time: Some methods, like bootstrap intervals, can be computationally intensive.
17. Interval instability: Intervals may vary significantly across samples.
18. Lack of uniqueness: Multiple intervals may be constructed for the same parameter.
19. Difficulty in hypothesis testing: Interval estimation may not directly facilitate hypothesis testing.
20. Limited generalizability: Intervals may not generalize well to other populations or contexts.

14.16 ERRORS IN INTERVAL ESTIMATION

Common errors in interval estimation:

1. Confidence interval misinterpretation: Misunderstanding the probability statement.
2. Incorrect confidence level: Using an inappropriate confidence level.
3. Assumption violations: Failing to check assumptions (e.g., normality, independence).
4. Sample size issues: Using too small or too large a sample size.

5. Outlier effects: Failing to account for outliers or influential observations.
6. Model misspecification: Using an incorrect or oversimplified model.
7. Non-coverage: Interval fails to contain the true parameter value.
8. Interval construction errors: Mistakes in calculating interval limits.
9. Units of measurement: Incorrect units or conversions.
10. Rounding errors: Rounding errors in calculations.
11. Software errors: Bugs or incorrect software usage.
12. Data entry errors: Incorrect data entry or transcription.
13. Failure to account for correlations: Ignoring correlations between variables.
14. Inappropriate interval method: Choosing the wrong interval estimation method.
15. Ignoring clustering or stratification: Failing to account for complex survey designs.
16. Not addressing missing data: Ignoring or improperly handling missing values.
17. Incorrect prior distributions: Specifying incorrect prior distributions in Bayesian intervals.
18. Failure to check for convergence: Not verifying convergence in iterative methods.
19. Ignoring model uncertainty: Failing to account for model uncertainty.
20. Lack of robustness checks: Not verifying interval robustness.

14.17 USES OF INTERVAL ESTIMATION

Interval estimation has numerous uses in various fields:

1. Medical Research: Estimate treatment effects, disease prevalence, and confidence intervals for survival rates.
2. Social Sciences: Estimate population means, proportions, and regression coefficients.
3. Engineering: Estimate system reliability, signal processing parameters, and confidence intervals for predictions.
4. Economics: Estimate economic indicators, forecast intervals, and confidence intervals for regression coefficients.
5. Quality Control: Monitor and control processes using tolerance intervals.
6. Environmental Studies: Estimate population sizes, growth rates, and confidence intervals for environmental indicators.
7. Pharmacokinetics: Estimate drug concentrations, absorption rates, and confidence intervals for pharmacokinetic parameters.
8. Image Processing: Estimate image parameters, confidence intervals for image segmentation.
9. Signal Processing: Estimate signal parameters, confidence intervals for signal detection.
10. Machine Learning: Estimate model parameters, confidence intervals for predictions.

11. Survey Research: Estimate population proportions, means, and confidence intervals.
12. Financial Analysis: Estimate stock prices, portfolio risk, and confidence intervals for financial forecasts.
13. Reliability Engineering: Estimate failure rates, confidence intervals for reliability metrics.
14. Marketing Research: Estimate market share, customer satisfaction, and confidence intervals.
15. Sports Analytics: Estimate team performance, player ratings, and confidence intervals.
16. Weather Forecasting: Estimate temperature ranges, precipitation confidence intervals.
17. Resource Allocation: Estimate resource requirements, confidence intervals.
18. Policy Evaluation: Estimate policy effects, confidence intervals.
19. Clinical Trials: Estimate treatment efficacy, confidence intervals.
20. Data Mining: Estimate patterns, relationships, and confidence intervals.

14.18 SUMMARY

Statistical estimation is the process of using sample data to make informed guesses or estimates about a population parameter. It involves: Pint Estimation and Interval Estimations

Estimating a single value for a parameter. Point estimation is a statistical method that involves estimating a single value for a population parameter based on a sample of data. The goal is to provide a best guess or a single estimate for the parameter.

Estimating a range of values for a parameter. Point estimation is a statistical method that involves estimating a single value for a population parameter based on a sample of data. The goal is to provide a best guess or a single estimate for the parameter.

14.19 TEST YOUR KNOWLEDGE

1. What is Point estimation ?
2. Explain the different Types of point estimators
3. What are the Properties of point estimator ?
4. Elaborate in your words three Point estimation methods
5. Highlight the Advantages of Point estimation?
6. What are the Disadvantages of Point estimation?
7. What are common Errors in Point Estimation?
8. Explain the uses of Point Estimation
9. What is Interval estimation?
10. Explain the different types of Interval estimation
11. What are the properties of Interval estimation?
12. Discuss the Interval estimation methods

13. What are the advantages of Interval estimation
14. Highlight the disadvantages of Interval estimation
15. What are the errors in Interval estimation?
16. Explain the uses of Interval estimation

14.20 FURTHER READINGS

1. Sankalp Gaurav, Business Statistics, Agra book International , Agra
2. Sinha, V.C., Principles of Statistics.
3. Gupta, S.B., Principles of Statistics.
4. Monga, G.S., Elementary Statistics.

UNIT 15 SAMPLING TEST

UNIT STRUCTURE

15.0 OBJECTIVE: After going **through** this unit you are able to understand

- Sampling and its methods
- Sampling test and hypothesis

15.1 Introduction

A sampling test involves selecting a subset of data (sample) from a larger population to make inferences about the population.

15.2 TYPES OF SAMPLING: sampling can be divided into two probability and non probability sampling.

15.2 1. Probability Sampling:

- Random Sampling: Every item has an equal chance.
- Stratified Sampling: Divide population into subgroups.
- Cluster Sampling: Select groups of items.

15.2 2. Non-Probability Sampling:

- Convenience Sampling: Easy-to-reach items.
- Purposive Sampling: Select items based on specific criteria.
- Quota Sampling: Fill predetermined quotas.

15.3 SAMPLING METHODS:

1**. Simple Random Sampling (SRS):** Simple Random Sampling is a probability sampling method where every member of the population has an equal chance of being selected. Each selection is independent, and the sample is drawn randomly without replacement.

2. **Systematic Sampling**: Systematic Sampling is a probability sampling method where every

nth member of the population is selected after a random start, providing a representative sample.

3. **Stratified Random Sampling**: Stratified Random Sampling is a probability sampling method that divides the population into distinct subgroups (strata) and then uses Simple Random Sampling within each stratum to select samples.

4. **Cluster Random Sampling**: Cluster Random Sampling is a probability sampling method where the population is divided into clusters (groups or areas), and a random selection of these clusters is chosen. All members within the selected clusters are included in the sample.

15.4 SAMPLING TEST STATISTICS:

1. Mean: he Mean, also known as the Average, is a statistical measure that calculates the central tendency of a dataset.

There are three types of Mean known as

1. Arithmetic Mean (AM): Most commonly used.

2. Geometric Mean (GM): Used for skewed distributions.

3. Harmonic Mean (HM): Used for rates and ratios.

2. Proportion: Proportion is a statistical measure that expresses the relationship between a part and the whole as a fraction or percentage.

There are three types of Proportions:

1. Simple Proportion: A fraction or percentage (e.g., 0.25 or 25%).

2. Relative Proportion: Compares two proportions (e.g., odds ratio).

3. Conditional Proportion: Depends on a specific condition (e.g., conditional probability).

3. Standard Deviation: Standard Deviation is the square root of the variance, representing how spread out values are from the mean.

Formula:

$SD = \sqrt{[(\Sigma(x - \mu)^2) / (n - 1)]}$

Where:

SD = Standard Deviation

Σ = Summation

x = Individual data points

μ = Mean

n = Sample size

4. Confidence Interval: A Confidence Interval (CI) estimates a population parameter (e.g., mean, proportion) with a specified level of confidence.

15.5 HYPOTHESIS TESTING:

1. Null Hypothesis (H0): The Null Hypothesis (H0) is a statement of no effect, no difference, or

no relationship between variables. It serves as a default assumption that there is no significant association or relationship between the variables being tested.

2. Alternative Hypothesis (H1): The Alternative Hypothesis (H1) is the opposite of the Null Hypothesis. It states that there is an effect, difference, or relationship.

3. Test Statistic: A Test Statistic is a numerical value calculated from sample data that helps determine whether to reject or fail to reject the Null Hypothesis (H0). It measures the difference between observed data and expected values under the Null Hypothesis.

4. P-Value: The P-Value, or Probability Value, is a statistical measure that represents the strength of evidence against the Null Hypothesis (H0). It's the probability of observing a result as extreme or more extreme than the one observed, assuming H0 is true.

15.6 Common Sampling Tests: there are mainly four types of sampling test as follows

1. One-Sample Z-Test

2. Two-Sample T-Test

3. ANOVA (Analysis of Variance)

4. Regression Analysis

Lets understand it one by one

15.6.1. One-Sample Z-Test: The One-Sample Z-Test is a statistical hypothesis test used to determine if a sample mean is significantly different from a known population mean.

Test Assumptions of One-Sample Z-Test:

1. Normality: Sample data follows a normal distribution.

2. Randomness: Sample is randomly selected.

3. Independence: Observations are independent.

4. Large Sample Size: $n \geq 30$ (or $n \geq 10$ with normality).

Test Statistics of One-Sample Z-Test:

1. Sample Mean ($\bar{x}$)

2. Population Mean (μ)

3. Sample Standard Deviation (s)

4. Standard Error ($SE = s / \sqrt{n}$)

5. Z-Score ($z = (\bar{x} - \mu) / SE$)

Hypothesis of One-Sample Z-Test:

1. Null Hypothesis (H0): $\bar{x} = \mu$

2. Alternative Hypothesis (H1): $\bar{x} \neq \mu$ (two-tailed), $\bar{x} > \mu$ (one-tailed), or $\bar{x} < \mu$ (one-tailed)

Test Procedure of One-Sample Z-Test:

1. Calculate sample mean and standard deviation.

2. Calculate standard error.

3. Calculate Z-score.

4. Determine critical Z-score (Zα/2) or P-value.

5. Compare Z-score to critical Z-score or P-value to α.

Interpretation of One-Sample Z-Test:

1. Reject H0 if Z-score > critical Z-score or P-value < α.

2. Fail to reject H0 if Z-score < critical Z-score or P-value > α.

Example of One-Sample Z-Test:

Suppose we want to test if the average height of a sample of 36 males is different from the known population mean of 175 cm.

$\bar{x}$ = 180 cm, s = 5 cm, μ = 175 cm, n = 36

SE = 5 / √36 = 0.83

z = (180 - 175) / 0.83 = 6.02

Using a Z-table or calculator, we find:

P-value = 0.0002 (two-tailed)

Since P-value < α (0.05), we reject H0.

2. Two-Sample T-Test: The Two-Sample T-Test compares the means of two independent groups to determine if there's a significant difference between them.

Test Assumptions of Two-Sample T-Test:

1. Normality: Both samples follow a normal distribution.

2. Independence: Observations are independent.

3. Equal Variances: Homogeneous variance across groups.

4. Randomness: Samples are randomly selected.

Test Statistics of Two-Sample T-Test:

1. Sample Mean 1 ($\bar{x}1$) and Sample Mean 2 ($\bar{x}2$)

2. Sample Standard Deviation 1 (s1) and Sample Standard Deviation 2 (s2)

3. Degrees of Freedom (df = n1 + n2 - 2)

4. T-Statistic (t = ($\bar{x}1$ - $\bar{x}2$) / SE)

5. Standard Error (SE = sqrt((s1^2/n1 + s2^2/n2))

Hypothesis of Two-Sample T-Test:

1. Null Hypothesis (H0): μ1 = μ2

2. Alternative Hypothesis (H1): μ1 ≠ μ2 (two-tailed), μ1 > μ2 (one-tailed), or μ1 < μ2 (one-

tailed)

Test Procedure of Two-Sample T-Test:

1. Calculate sample means and standard deviations.

2. Calculate degrees of freedom.

3. Calculate T-statistic.

4. Determine critical T-score (Tα/2) or P-value.

5. Compare T-statistic to critical T-score or P-value to α.

Interpretation of Two-Sample T-Test:

1. Reject H0 if T-statistic > critical T-score or P-value < α.

2. Fail to reject H0 if T-statistic < critical T-score or P-value > α.

Example of Two-Sample T-Test:

Suppose we compare exam scores of students taught by Method A (n1 = 25, $\bar{x}1$ = 85) and Method B (n2 = 30, $\bar{x}2$ = 90).

s1 = 10, s2 = 12

t = (85 - 90) / sqrt((10^2/25 + 12^2/30)) = -2.34

df = 25 + 30 - 2 = 53

Using a T-table or calculator, we find:

P-value = 0.022 (two-tailed)

Since P-value < α (0.05), we reject H0.

3. ANOVA (Analysis of Variance): ANOVA is a statistical technique used to compare means of three or more groups to determine if there's a significant difference between them.

Types of ANOVA:

1. One-Way ANOVA: Compare means across three or more groups.

2. Two-Way ANOVA: Examine interactions between two factors.

3. Repeated Measures ANOVA: Compare means over time or conditions.

Assumptions of ANOVA:

1. Normality: Data follows a normal distribution.

2. Independence: Observations are independent.

3. Homogeneity of Variances: Equal variances across groups.

4. Randomness: Samples are randomly selected.

ANOVA Hypothesis:

1. Null Hypothesis (H0): μ1 = μ2 = ... = μk (means are equal).

2. Alternative Hypothesis (H1): Not all means are equal.

ANOVA Statistics:

1. F-Ratio (F = MSB / MSW)

2. Mean Square Between (MSB)

3. Mean Square Within (MSW)

4. Degrees of Freedom (dfB, dfW)

5. P-Value

ANOVA Procedure:

1. Calculate group means and variances.

2. Calculate MSB and MSW.

3. Calculate F-Ratio.

4. Determine critical F-Score or P-Value.

5. Compare F-Ratio to critical F-Score or P-Value to α.

ANOVA Interpretation:

1. Reject H0 if F-Ratio > critical F-Score or P-Value < α.

2. Fail to reject H0 if F-Ratio < critical F-Score or P-Value > α.

4. Regression Analysis: Regression analysis is a statistical method used to establish a relationship between two or more variables. It helps us understand how a dependent variable (outcome or response) changes when one or more independent variables (predictors or explanatory variables) are varied.

Types of Regression Analysis:

1. Simple Linear Regression: One independent variable is used to predict the dependent variable.

2. Multiple Linear Regressions: Two or more independent variables are used to predict the dependent variable.

3. Non-Linear Regression: Relationships between variables are modelled using non-linear functions (e.g., polynomial, exponential).

4. Logistic Regression: Used for binary classification problems (0/1, yes/no).

5. Ridge Regression: Regularized regression to reduce multi collinearity.

6. Lasso Regression: Regularized regression to select relevant features.

7. Elastic Net Regression: Combines Ridge and Lasso regression.

Key Concepts of Regression Analysis:

1. Coefficient of Determination (R-squared): Measures the goodness of fit.

2. Coefficient of Correlation: Measures the strength and direction of the relationship.

3. P-value: Indicates the significance of the relationship.

4. Confidence Interval: Provides a range of plausible values for the regression coefficients.

5. Residuals: Differences between observed and predicted values.

Assumptions of Regression Analysis:

1. Linearity: Relationship between variables is linear.

2. Independence: Observations are independent.

3. Homoscedasticity: Constant variance of residuals.

4. Normality: Residuals follow a normal distribution.

5. No multicollinearity: Independent variables are not highly correlated.

15.7 Summary

A sampling test involves selecting a subset of data (sample) from a larger population to make inferences about the population.

15.8 Test Your Knowledge

1. Explain the various types of Sampling

.....
.....
.....

11. What are different Sampling Methods? Discuss

.....
.....
.....
.....

3. What is Sampling Test Statistics?

.....
.....
.....
.....

4. What is Hypothesis Testing?

.....
.....
.....
.....

5. Discuss Common Sampling Tests

.....
.....
.....
.....

15. 9 further readings

1. Sankalp Gaurav, Business Statistics, Agra book International , Agra
2. Sinha, V.C., Principles of Statistics.
3. Gupta, S.B., Principles of Statistics.
4. Monga, G.S., Elementary Statistics.

UNIT – 16 HYPOTHESIS TESTING

UNIT STRUCTURE:

16.0 OBJECTIVES :

After going through this unit you should be able to know about the–

1. Conceptual Framework of Hypothesis
2. Types of hypothesis and testing of hypothesis
3. Hypothesis Errors

16.1. INTRODUCTION OR CONCEPTUAL FRAMEWORK OF HYPOTHESIS

A hypothesis (plural *hypotheses*) is a proposed explanation for a phenomenon. For a hypothesis to be a scientific hypothesis, the scientific method requires that one can test it. Scientists generally base scientific hypotheses on previous observations that cannot satisfactorily be explained with the available scientific theories. Even though the words "hypothesis" and "theory" are often used synonymously, a scientific hypothesis is not the same as a scientific theory. A working hypothesis is a provisionally accepted hypothesis proposed for further research.

A different meaning of the term hypothesis is used in formal logic, to denote the antecedent of a proposition; thus in the proposition "If P, then Q". P denotes the hypothesis (or antecedent); Q can be called a consequent. P is the assumption in a (possibly counterfactual) what if question.

The adjective hypothetical, meaning "having the nature of a hypothesis", or "being assumed to exist as an immediate consequence of a hypothesis", can refer to any of these meanings of the term "hypothesis".

16.2. USES OF HYPOTHESIS

In its ancient usage, hypothesis referred to a summary of the plot of a classical drama the english word hypothesis comes from the ancient Greek ύπόθεσις (hupothesis), meaning "to put under" or "to suppose".

In Plato's Meno (86e-87b), Socrates dissects virtue with a method used by mathematicians, that of "investigating from a hypothesis". In this sense, 'hypothesis' refers to a clever idea or to a

convenient mathematical approach that simplifies cumbersome calculations. Cardinal Bellarmine gave a famous example of this usage in the warning issued to Galileo in the early 17th century that he must not treat the motion of the earth as a reality, but merely as a hypothesis.

In common usage in the 21st century. a hypothesis refers to a provisional idea whose merit requires evaluation. For proper evaluation, the framer of a hypothesis needs to define specifics in operational terms. A hypothesis requires more work by the researcher in order to either confirm or disprove it in due course, a confirmed hypothesis may become part of a theory or occasionally may grew to become a theory itself. Normally, scientific hypotheses have the form of a mathematical model. Sometimes, but not always, one can also formulate them as existential statements, stating that some particular instance of the phenomenon under examination has some characteristic and causal explanations, which have the general form of universal statements, stating that every instance of the phenomenon has a particular characteristic.

Any useful hypothesis will enable predictions by reasoning (including deductive reasoning). It might predict the outcome of an experiment in a laboratory setting or the observation of a phenomenon in nature. The prediction may also invoke statistics and only talk about probabilities. Karl Popper, following others, has argued that a hypothesis must be falsifiable, and that one cannot regard a proposition or theory as scientific if it does not admit the possibility of being shown false. Other philosophers of science have rejected the criterion of falsifiability or supplemented it with other criteria, such as verifiability (e.g., verificationism) or coherence (e.g., confirmation holism). The scientific method involves experimentation, to test the ability of some hypothesis to adequately answer the question under investigation. In contrast, unfettered observation is not as likely to raise unexplained issues or open questions in science, as would the formulation of a crucial experiment to test the hypothesis. A thought experiment might also be used to test the hypothesis as well.

In framing a hypothesis, the investigator must not currently know the outcome of a test or that it remains reasonably under continuing investigation. Only in such cases does the experiment, test or study potentially increase the probability of showing the truth of a hypothesis. If the researcher already knows the outcome, it counts as a "consequence" and the researcher should have already considered this while formulating the hypothesis. If one cannot assess the predictions by observation or by experience, the hypothesis needs to be tested by others providing observations. For example, a new technology or theory might make the necessary experiments feasible.

16.3. SCIENTIFIC HYPOTHESIS

People refer to a trial solution to a problem as a hypothesis, often called an "educated guess" because it provides a suggested solution based on the evidence. However, some scientists reject the term "educated guess" as incorrect. Experimenters may test and reject several hypotheses before solving the problem.

Working Hypothesis

A working hypothesis is a hypothesis that is provisionally accepted as a basis for further research in the hope that a tenable theory will be produced, even if the hypothesis ultimately fails. Like all hypotheses, a working hypothesis is constructed as a statement of expectations, which can be linked to the exploratory research purpose in empirical investigation. Working hypotheses and are often used as a conceptual framework in qualitative research.

The provisional nature of working hypotheses make them useful as an organizing device in applied research. Here they act like a useful guide to address problems that are still in a formative phase.

In recent years, philosophers of science have tried to integrate the various approaches to evaluating

hypotheses, and the scientific method in general, to form a more complete system that integrates the individual concerns of each approach. Notably, Imre Lakatos and Paul Feyerabend, Karl Popper's colleague and student, respectively, have produced novel attempts at such a synthesis.

16.4. MEASUREMENT OF HYPOTHESIS

Concepts in Hempel's deductive homological model play key role in the development and testing of hypothesis. Most formal hypothesis connect concepts by specifying the expected relationships between propositions. When a set of hypothesis are grouped together they become a type of conceptual framework. When a conceptual framework is complex and incorporates causality or explanation it is generally referred to as a theory. According to noted philosopher of science Carl Gustav Hempel " an adequate empirical interpretation turns a theoretical system into a testable theory. The hypothesis whose constituent terms have been interpreted become capable of test by reference to observable phenomena. Frequently the interpreted hypothesis will be derivative hypotheses of the theory; but their confirmation or disconfirmation by empirical data will then immediately strengthen or weaken also the primitive hypotheses from which they were derived."

Hempel provides a useful metaphor that describes the relationship between a conceptual framework and the framework as it is observed and perhaps tested (interpreted framework). "The whole system floats, as it were, above the plane of observation and is anchored to it by rules of interpretation. These might be viewed as strings which are not part of the network but link certain points of the latter with specific places in the plane of observation. By virtue of those interpretative connections, the network can function as a scientific theory". Hypotheses with concepts anchored in the plane of observation are ready to be tested. In "actual scientific practice the process of framing a theoretical structure and of interpreting it are not always sharply separated, since the intended interpretation usually guides the construction of the theoretician". It is, however, "possible and indeed desirable, for the purposes of logical clarification, to separate the two steps conceptually."

16.5. STATISTICAL HYPOTHESIS TESTING

When a possible correlation or similar relation between phenomena is investigated, such as whether a proposed remedy is effective in treating a disease, the hypothesis that a relation exists cannot be examined the same way one might examine a proposed new law of nature. In such an investigation, if the tested remedy shows no effect in a few cases, these do not necessarily falsify the hypothesis. Instead, statistical tests are used to determine how likely it is that the overall effect would be observed if the hypothesized relation does not exist. If that likelihood is sufficiently small (e.g., less than 1%), the existence of a relation may be assumed. Otherwise, any observed effect may be due to pure chance.

In statistical hypothesis testing, two hypotheses are compared. These are called the null hypothesis and the alternative hypothesis. The null hypothesis is the hypothesis that states that there is no relation between the phenomena whose relation is under investigation, or at least not of the form given by the alternative hypothesis. The alternative hypothesis, as the name suggests, is the alternative to the null hypothesis. it states that there is some kind of relation. The alternative hypothesis may take several forms, depending on the nature of the hypothesized relation; in particular it can be two-sided (for example: there is some effect, in a yet unknown direction) or one-sided (the direction of the hypothesized relation positive or negative, is fixed in advance).

Conventional significance levels for testing hypotheses (acceptable probabilities of wrongly rejecting a true null hypothesis) are .10, .05, and .01. Whether the null hypothesis is rejected and the alternative hypothesis is accepted, must be determined in advance, before the observations are collected or inspected. If these criteria are determined later, when the data to be tested are already

known, the test is invalid.

The above procedure is actually dependent on the number of the participants (units or sample size) that is included in the study. For instance, the sample size may be too small to reject a null hypothesis and, therefore, it is recommended to specify the sample size from the beginning. It is advisable to define a small, medium and large effect size for each of a lumber of important statistical tests which are used to test the hypotheses.

16.6 HYPOTHESIS ERROR

Using hypothesis testing, you can make decisions about whether your data support or refute your research predictions with null and alternative hypotheses.

Hypothesis testing starts with the assumption of no difference between groups or no relationship between variables in the population—this is the **null hypothesis**. It's always paired with an **alternative hypothesis**, which is your research prediction of an actual difference between groups or a true relationship between variables.

Then, you decide whether the null hypothesis can be rejected based on your data and the results of a statistical test. Since these decisions are based on probabilities, there is always a risk of making the wrong conclusion.

If your results show statistical significance, that means they are very unlikely to occur if the null hypothesis is true. In this case, you would reject your null hypothesis. But sometimes, this may actually be a Type I error.

If your findings do not show statistical significance, they have a high chance of occurring if the null hypothesis is true. Therefore, you fail to reject your null hypothesis. But sometimes, this may be a Type II error.

16.7 TYPE I ERROR

A Type I error means rejecting the null hypothesis when it's actually true. It means concluding that results are **statistically significant** when, in reality, they came about purely by chance or because of unrelated factors.

The risk of committing this error is the significance level (alpha or α) you choose. That's a value that you set at the beginning of your study to assess the statistical probability of obtaining your results (p value).

The significance level is usually set at 0.05 or 5%. This means that your results only have a 5% chance of occurring, or less, if the null hypothesis is actually true.

If the p value of your test is lower than the significance level, it means your results are statistically significant and consistent with the alternative hypothesis. If your p value is higher than the significance level, then your results are considered statistically non-significant.

16.8 Type II Error

A Type II error means not rejecting the null hypothesis when it's actually false. This is not quite the same as "accepting" the null hypothesis, because hypothesis testing can only tell you whether to reject the null hypothesis.

Instead, a Type II error means failing to conclude there was an effect when there actually was. In reality, your study may not have had enough **statistical power** to detect an effect of a certain

size.

Power is the extent to which a test can correctly detect a real effect when there is one. A power level of 80% or higher is usually considered acceptable.

The risk of a Type II error is inversely related to the statistical power of a study. The higher the statistical power, the lower the probability of making a Type II error.

16.9 SUMMARY

A hypothesis (plural *hypotheses*) is a proposed explanation for a phenomenon. For a hypothesis to be a scientific hypothesis, the scientific method requires that one can test it. Scientists generally base scientific hypotheses on previous observations that cannot satisfactorily be explained with the available scientific theories.

In its ancient usage, hypothesis referred to a summary of the plot of a classical drama the english word hypothesis comes from the ancient Greek ύπόθεσις (hupothesis), meaning "to put under" or "to suppose". People refer to a trial solution to a problem as a hypothesis, often called an "educated guess" because it provides a suggested solution based on the evidence. However, some scientists reject the term "educated guess" as incorrect. Experimenters may test and reject several hypotheses before solving the problem.

A working hypothesis is a hypothesis that is provisionally accepted as a basis for further research in the hope that a tenable theory will be produced, even if the hypothesis ultimately fails. Like all hypotheses, a working hypothesis is constructed as a statement of expectations, which can be linked to the exploratory research purpose in empirical investigation. Working hypotheses and are often used as a conceptual framework in qualitative research.

16.10 TEST YOUR KNOWLEDGE

1. What is Hypothesis?
2. What are the two types of errors?
3. Explain type I error
4. Elaborate the conditions in which type II error occurs?
5. What is hypothesis error?
6. Explain working hypothesis
7. Define scientific hypothesis?

16.11 FURTHER READINGS

1. Sankalp Gaurav, Business Statistics, Agra book International , Agra
2. Sinha, V.C., Principles of Statistics.
3. Gupta, S.B., Principles of Statistics.
4. Monga, G.S., Elementary Statistics.

UNIT 17 LARGE & SMALL SAMPLES

UNIT STRUCTURE

17.0 OBJECTIVE: After going through this unit you are able to understand

- Large sample and small sample
- Advantages and limitations of large and small sample
- Types of small and large sample test

17.1 LARGE SAMPLE

In statistics, a large sample refers to a sample size that is sufficiently big to accurately represent the population from which it was drawn. The exact definition of "large" can vary depending on the context and the statistical analysis being performed, but here are some general guidelines:

1. Sample size greater than 30: In many statistical tests, a sample size of 30 or more is considered large enough to assume normality and apply parametric tests.
2. Sample size greater than 100: For more precise estimates and narrower confidence intervals, a sample size of 100 or more is often recommended.
3. Sample size at least 10% of the population: If the population size is known, a large sample is often defined as at least 10% of the population.

17.2 BENEFITS OF LARGE SAMPLE

The benefits of a large sample include:

1. Increased Precision: Larger samples provide more accurate estimates of population parameters, reducing margin of error.
2. Better Representation: Large samples are more likely to represent the population's diversity, reducing bias.
3. Improved Reliability: Results from large samples are more reliable and generalizable to the population.
4. Increased Statistical Power: Large samples can detect smaller effects, differences, and relationships.
5. Reduced Sampling Error: Larger samples minimize sampling error, ensuring results are closer to population values.
6. More Stable Estimates: Large samples provide more stable estimates of population parameters.
7. Enhanced Generalizability: Results from large samples can be applied to the broader population.
8. Increased Confidence: Large samples boost confidence in findings, supporting informed decision-making.
9. Better Detection of Relationships: Large samples help identify complex relationships and interactions.
10. Improved Subgroup Analysis: Large samples enable analysis of smaller subgroups within the population.
11. Reduced Variance: Larger samples reduce variance, leading to more consistent results.
12. More Accurate Hypothesis Testing: Large samples enhance hypothesis testing, reducing Type I and Type II errors.
13. Better Handling of Non-Response: Large samples mitigate the impact of non-response or missing data.
14. Enhanced Data Modeling: Large samples support more complex data modeling and analysis.
15. Increased Credibility: Studies with large samples are often viewed as more credible and authoritative.

17.3 LIMITATION OF LARGE SAMPLE

While large samples offer many benefits, they also come with some limitations:

1. Increased Cost: Collecting and analyzing large samples can be expensive and resource-intensive.
2. Data Quality Issues: Larger samples increase the risk of data entry errors, missing values, and inconsistencies.
3. Analysis Complexity: Large samples require advanced statistical techniques and computational power.
4. Over fitting: Large samples can lead to over fitting, where models become too complex and

perform poorly on new data.

5. Diminishing Returns: Beyond a certain point, increasing sample size yields minimal gains in precision.
6. Ethical Concerns: Large samples may raise ethical concerns, such as privacy violations or participant burden.
7. Sampling Bias: Large samples can still suffer from sampling bias if selection methods are flawed.
8. Data Management Challenges: Storing, managing, and analyzing large datasets can be cumbersome.
9. Increased Risk of False Positives: Large samples increase the likelihood of detecting statistically significant but practically insignificant effects.
10. Computational Time and Resources: Analyzing large samples requires significant computational power and time.
11. Model Assumptions: Large samples can mask violations of model assumptions, leading to inaccurate conclusions.
12. Interpretation Challenges: Large samples can make it difficult to interpret results, especially for complex models.
13. Data Privacy and Security: Large samples increase the risk of data breaches and privacy violations.
14. Participant Fatigue: Large samples can lead to participant fatigue, decreasing response rates and data quality.
15. Environmental Impact: Large-scale data collection can have a significant environmental impact.

17.4 TEST FOR LARGE SAMPLE

Here are some common tests used for large samples:

1. Z-tests: For means, proportions, and differences between means and proportions.

2. T-tests: For means and differences between means (when population standard deviation is unknown).

3. Chi-Square tests: For categorical data, goodness of fit, and independence.

4. ANOVA (Analysis of Variance): For comparing means across multiple groups.

5. Regression Analysis: For modeling relationships between variables.

6. Confidence Intervals: For estimating population parameters.

7. Non-parametric tests (e.g., Wilcoxon, Kruskal-Wallis): For non-normal data or ordinal data.

8. Time Series Analysis: For analyzing data with temporal dependencies.

9. Survey Sampling Methods: For analyzing data from complex survey designs.

10. Bootstrap Methods: For estimating standard errors and confidence intervals.

17.5 TESTS FOR LARGE SAMPLES INCLUDE:

17.5.1. Large Sample Test of Proportions: The Large Sample Test of Proportions is a statistical test used to compare the proportion of successes (or occurrences) in two or more independent groups. It's also known as the Two-Proportion Z-Test or the Large Sample Test for Differences in Proportions.

Assumptions:

1. Independent samples

2. Large sample sizes ($n1 \geq 30$, $n2 \geq 30$)

3. Binary response variable (success/failure, yes/no, etc.)

4. Random sampling

Test Statistic:

$Z = (p1 - p2) / \text{sqrt}(p*(1-p)*(1/n1 + 1/n2))$

where:

p1, p2 = sample proportions

p = pooled proportion = $(x1 + x2) / (n1 + n2)$

x1, x2 = number of successes in each group

n1, n2 = sample sizes

Null and Alternative Hypotheses:

H0: $p1 = p2$ (no difference in proportions)

H1: $p1 \neq p2$ (difference in proportions)

Decision Rule:

Compare the calculated Z-score to the critical Z-score from the standard normal distribution (Z-table) or use a p-value.

Interpretation:

- If $Z \geq Z\alpha/2$ or p-value $\leq \alpha$, reject H0 (significant difference in proportions).

- Otherwise, fail to reject H0 (no significant difference).

Example:

Suppose we want to compare the proportion of smokers in two cities:

City A: 120 smokers out of 1000 ($p1 = 0.12$)

City B: 150 smokers out of 1200 ($p2 = 0.125$)

Using the Large Sample Test of Proportions, we calculate:

$Z = (0.12 - 0.125) / \text{sqrt}(0.12375*(1-0.12375)*(1/1000 + 1/1200)) \approx -0.43$

With $\alpha = 0.05$, the critical Z-score is ± 1.96. Since $-0.43 < -1.96$, we fail to reject H0, indicating

no significant difference in smoking proportions between the two cities.

This test helps determine if the difference in proportions is statistically significant, allowing informed decisions in various fields like medicine, social sciences, and marketing.

17.1.2. Large Sample Test of Means: The Large Sample Test of Means, also known as the Z-Test for Means, is a statistical test used to compare the means of two independent groups. It's suitable for large samples ($n \geq 30$).

Assumptions:

1. Independent samples

2. Large sample sizes ($n1 \geq 30$, $n2 \geq 30$)

3. Random sampling

4. Normality or approximately normal distribution

5. Equal variances (homogeneity of variance)

Test Statistic:

$Z = (\bar{x}1 - \bar{x}2) / \text{sqrt}((\sigma1^2 / n1) + (\sigma2^2 / n2))$

where:

$\bar{x}1$, $\bar{x}2$ = sample means

$\sigma1$, $\sigma2$ = population standard deviations

n1, n2 = sample sizes

Null and Alternative Hypotheses:

H0: $\mu1 = \mu2$ (no difference in means)

H1: $\mu1 \neq \mu2$ (difference in means)

Decision Rule:

Compare the calculated Z-score to the critical Z-score from the standard normal distribution (Z-table) or use a p-value.

Interpretation:

- If $Z \geq Z\alpha/2$ or p-value $\leq \alpha$, reject H0 (significant difference in means).

- Otherwise, fail to reject H0 (no significant difference).

Example:

Suppose we want to compare the average heights of men and women:

Men: $\bar{x}1 = 175.5$ cm, $\sigma1 = 5.2$ cm, $n1 = 50$

Women: $\bar{x}2 = 162.8$ cm, $\sigma2 = 4.8$ cm, $n2 = 60$

Using the Large Sample Test of Means, we calculate:

Z = (175.5 - 162.8) / sqrt((5.2^2 / 50) + (4.8^2 / 60)) ≈ 4.23

With α = 0.05, the critical Z-score is ±1.96. Since 4.23 > 1.96, we reject H0, indicating a significant difference in average heights between men and women.

Variations:

- One-sample Z-Test: Compare a sample mean to a known population mean.

- Paired Z-Test: Compare means from paired or matched data.

This test helps determine if the difference in means is statistically significant, commonly used in fields like medicine, social sciences, and business.

17.5.3. Large Sample Confidence Interval for Proportions: The Large Sample Confidence Interval for Proportions is a statistical method to estimate a population proportion (p) based on a sample proportion ($\hat{p}$). It's used when:

1. Sample size (n) is large ($n \geq 30$)

2. Sample proportion ($\hat{p}$) is not close to 0 or 1

Formula:

$\hat{p} \pm Z\alpha/2 * \sqrt{(\hat{p}(1-\hat{p})/n)}$

where:

- $\hat{p}$ = sample proportion

- $Z\alpha/2$ = critical value from standard normal distribution (Z-table)

- n = sample size

- α = significance level (e.g., 0.05 for 95% CI)

Steps:

1. Calculate sample proportion ($\hat{p}$)

2. Choose significance level (α) and find $Z\alpha/2$

3. Plug values into formula

4. Calculate margin of error (ME) = $Z\alpha/2 * \sqrt{(\hat{p}(1-\hat{p})/n)}$

5. Calculate confidence interval (CI) = $\hat{p} \pm ME$

Example:

Suppose we want to estimate the proportion of smokers in a population.

Sample size (n) = 1000

Number of smokers = 120

Sample proportion ($\hat{p}$) = 120/1000 = 0.12

For 95% CI (α = 0.05), $Z\alpha/2 = 1.96$

ME = 1.96 * √(0.12(1-0.12)/1000) ≈ 0.026

CI = 0.12 ± 0.026 = (0.094, 0.146)

Interpretation:

We are 95% confident that the true population proportion of smokers lies between 0.094 and 0.146.

Note:

- Large sample size ensures the sampling distribution of $\hat{p}$ is approximately normal.

- If $\hat{p}$ is close to 0 or 1, consider alternative methods like exact binomial confidence intervals.

This confidence interval helps estimate population proportions, commonly used in surveys, medical research, and social sciences.

17.5.4. Large Sample Confidence Interval for Means: The Large Sample Confidence Interval for Means is a statistical method to estimate a population mean (μ) based on a sample mean ($\bar{x}$). It's used when:

1. Sample size (n) is large ($n \geq 30$)

2. Population standard deviation (σ) is known or estimated

Formula:

$\bar{x} \pm Z\alpha/2 * (\sigma / \sqrt{n})$

where:

- $\bar{x}$ = sample mean

- $Z\alpha/2$ = critical value from standard normal distribution (Z-table)

- σ = population standard deviation (or estimated standard deviation, s)

- n = sample size

- α = significance level (e.g., 0.05 for 95% CI)

Steps:

1. Calculate sample mean ($\bar{x}$)

2. Choose significance level (α) and find $Z\alpha/2$

3. Plug values into formula

4. Calculate margin of error (ME) = $Z\alpha/2 * (\sigma / \sqrt{n})$

5. Calculate confidence interval (CI) = $\bar{x} \pm ME$

Example:

Suppose we want to estimate the average height of adults.

Sample size (n) = 50

Sample mean ($\bar{x}$) = 175.5 cm

Population standard deviation (σ) = 5.2 cm

For 95% CI (α = 0.05), Zα/2 = 1.96

ME = 1.96 * (5.2 / √50) ≈ 1.43

CI = 175.5 ± 1.43 = (174.07, 176.93)

Interpretation:

We are 95% confident that the true population mean height lies between 174.07 cm and 176.93 cm.

Note:

- Large sample size ensures the sampling distribution of $\bar{x}$ is approximately normal.

- If σ is unknown, use the sample standard deviation (s) and a t-distribution for smaller samples.

This confidence interval helps estimate population means, commonly used in various fields like medicine, social sciences, and engineering.

17.5.5. Chi-Square Test for Goodness of Fit (large sample): The Chi-Square Test for Goodness of Fit is a statistical test used to determine how well a theoretical distribution (e.g., normal, binomial) fits a observed frequency distribution. It assesses whether the observed frequencies significantly differ from the expected frequencies based on the theoretical distribution.

Test Statistic:

$\chi^2 = \Sigma$ [(observed frequency - expected frequency)^2 / expected frequency]

where:

- χ^2 = Chi-Square statistic

- Σ = summation over all categories

- observed frequency = actual count in each category

- expected frequency = count predicted by theoretical distribution

Degrees of Freedom (df):

- df = number of categories - 1

Null Hypothesis (H0):

- The observed frequencies follow the theoretical distribution.

Alternative Hypothesis (H1):

- The observed frequencies do not follow the theoretical distribution.

Decision Rule:

- Calculate χ^2 and df

- Find critical χ^2 value or p-value from Chi-Square distribution

- If $\chi^2 \geq$ critical value or p-value $\leq \alpha$, reject H0 (significant difference)

- Otherwise, fail to reject H0 (no significant difference)

Example:

Suppose we want to test if a coin is fair (theoretical distribution: 50% heads, 50% tails).

Observed frequencies:

- Heads: 45

- Tails: 55

Expected frequencies (based on 50% probability):

- Heads: 50

- Tails: 50

$\chi^2 = [(45-50)^2/50 + (55-50)^2/50] = 1$

df = 2 - 1 = 1

Critical χ^2 value ($\alpha = 0.05$, df = 1) = 3.84

Since $\chi^2 < 3.84$, we fail to reject H0, indicating the coin is likely fair.

17.5.6. Kolmogorov-Smirnov Test (large sample): The Kolmogorov-Smirnov Test (K-S Test) is a non-parametric test used to determine if a sample comes from a specific distribution (e.g., normal, uniform). It's particularly useful for large samples.

Test Statistic:

$D = \sup|F(x) - G(x)|$

where:

- D = maximum distance between empirical distribution function (F(x)) and cumulative distribution function (G(x))

- F(x) = proportion of sample values $\leq$ x

- G(x) = cumulative distribution function of the hypothesized distribution

Null Hypothesis (H0):

- The sample comes from the specified distribution.

Alternative Hypothesis (H1):

- The sample does not come from the specified distribution.

Decision Rule:

- Calculate D

- Find critical value or p-value from K-S distribution

- If D $\geq$ critical value or p-value $\leq \alpha$, reject H0 (sample does not come from specified distribution)

- Otherwise, fail to reject H0 (sample likely comes from specified distribution)

Large Sample Approximation:

For large samples (n > 30), use the following approximation:

$D \approx \sqrt{(-0.5 * \ln(\alpha/2))} * \sqrt{(n)}$

where α is the significance level.

Example:

Suppose we want to test if a sample of 50 data points comes from a standard normal distribution.

$D = 0.15$

Using the large sample approximation:

$D \approx \sqrt{(-0.5 * \ln(0.05/2))} * \sqrt{(50)} \approx 0.173$

Since $D < 0.173$, we fail to reject H0, indicating the sample likely comes from a standard normal distribution.

Advantages:

- ✓ Non-parametric, no distributional assumptions
- ✓ Suitable for large samples
- ✓ Can test for any continuous distribution

Limitations:

- ✓ Not suitable for small samples
- ✓ Not sensitive to differences in distribution tails
- ✓ Can be conservative (failing to reject H0 when false)

The K-S Test is widely used in:

- ✓ Statistical hypothesis testing
- ✓ Distributional analysis
- ✓ Goodness-of-fit testing
- ✓ Quality control

17.6 SMALL SAMPLE

Small Sample refers to a statistical sample with a limited number of observations, typically fewer than 30. When dealing with small samples:

1. Central Limit Theorem (CLT) may not apply.

2. Distributional assumptions are crucial.

3. Statistical tests may have reduced power.

4. Confidence intervals may be wider.

17.7 Benefits of Small Sample

While small samples present challenges, they also offer benefits:

1. Cost-effective: Collecting and analyzing smaller datasets is often less expensive.

2. Time-efficient: Smaller samples require less time for data collection and analysis.

3. Resource-efficient: Small samples conserve resources, making them ideal for pilot studies or exploratory research.

4. Increased control: Smaller samples allow for more control over data quality and experimental conditions.

5. Personalized insights: Small samples provide detailed information about individual cases or subgroups.

6. Qualitative research: Small samples are well-suited for in-depth, qualitative analysis.

7. Hypothesis generation: Small samples help generate hypotheses for further investigation.

8. Proof-of-concept: Small samples demonstrate feasibility and potential for larger studies.

9. Reduced participant burden: Smaller samples minimize participant fatigue and burden.

10. Rapid iteration: Small samples enable quick testing and refinement of research questions or methods.

11. Enhanced data quality: Smaller samples allow for more thorough data cleaning and validation.

12. Contextual understanding: Small samples provide rich contextual information.

13. Innovation: Small samples facilitate innovative, exploratory research.

14. Collaboration: Smaller samples encourage collaboration among researchers.

15. Educational: Small samples serve as teaching tools for research methods and statistics.

While small samples have limitations, their benefits make them valuable in various research contexts, such as:

- Pilot studies
- Case studies
- Qualitative research
- Exploratory research
- Resource-constrained projects
- Proof-of-concept studies
- Innovation and development

17.8 LIMITATIONS OF SMALL SAMPLE

Small samples have several limitations:

1. Lack of representation: Small samples may not accurately represent the population.
2. Increased variability: Small samples are more susceptible to random fluctuations.
3. Reduced precision: Estimates and confidence intervals are less precise.
4. Limited generalizability: Results may not apply to larger populations.
5. Increased risk of bias: Small samples are more prone to selection bias.
6. Difficulty detecting effects: Small samples may fail to detect statistically significant effects.
7. Overemphasis on outliers: Small samples can be heavily influenced by extreme values.
8. Limited subgroup analysis: Small samples restrict subgroup analysis and exploration.
9. Insufficient power: Small samples often lack sufficient statistical power.
10. Difficulty estimating rare events: Small samples may not capture rare occurrences.
11. Model overfitting: Small samples increase the risk of overfitting in modeling.
12. Limited data quality: Small samples may not allow for thorough data cleaning.
13. Increased standard error: Small samples result in larger standard errors.
14. Reduced reliability: Small samples can lead to less reliable results.
15. Difficulty publishing: Small sample studies may face challenges getting published.

To mitigate these limitations:

1. Use appropriate statistical methods.
2. Clearly report limitations.
3. Consider bootstrap or simulation methods.
4. Focus on effect sizes rather than p-values.
5. Use Bayesian inference.
6. Collect more data if possible.
7. Use data augmentation techniques.
8. Consider meta-analysis.
9. Emphasize exploratory nature.
10. Interpret results cautiously.

17.9 TYPES OF SMALL SAMPLE

Small samples can be categorized into various types based on their characteristics:

1. Convenience Sample: Easily accessible participants.
2. Purposive Sample: Selected based on specific criteria.

3. Snowball Sample: Participants recruit additional participants.
4. Pilot Sample: Small-scale study preceding a larger study.
5. Case Study Sample: In-depth examination of a single case.
6. Expert Sample: Specialists or experts in a particular field.
7. Homogeneous Sample: Participants share similar characteristics.
8. Heterogeneous Sample: Participants have diverse characteristics.
9. Random Sample: Participants selected randomly.
10. Stratified Sample: Subgroups within the population.
11. Cluster Sample: Participants from specific groups or clusters.
12. Systematic Sample: Participants selected at regular intervals.
13. Judgment Sample: Selected based on researcher expertise.
14. Quota Sample: Participants selected to meet specific quotas.
15. Longitudinal Sample: Same participants studied over time.
16. Cross-Sectional Sample: Participants studied at a single point.
17. Panel Sample: Participants studied repeatedly.
18. Cohort Sample: Participants sharing a common experience.

These types of small samples serve different research purposes and offer unique advantages. Understanding the characteristics of each type helps researchers:

- Select appropriate sampling methods
- Ensure representativeness
- Minimize bias
- Maximize data quality
- Interpret results accurately

17.10 Tests for small samples:

1. t-test (Student's t-test): Compares means between two groups.
2. Wilcoxon Rank-Sum Test (Mann-Whitney U Test): Compares distributions between two groups.
3. Sign Test: Tests median or proportion.
4. Fisher's Exact Test: Tests association in 2x2 contingency tables.
5. Chi-Square Test (Goodness-of-Fit): Tests distribution fit.
6. Kolmogorov-Smirnov Test: Tests distribution equality.
7. ANOVA (Analysis of Variance): Compares means across multiple groups.

8. Kruskal-Wallis H Test: Compares distributions across multiple groups.

9. Friedman Test: Compares treatments across multiple blocks.

10. Exact Tests (e.g., Binomial, Poisson): Tests specific distributions.

11. Bootstrap Methods: Resamples data for estimation and testing.

12. Permutation Tests: Randomly rearranges data for testing.

13. Bayesian Methods: Incorporates prior knowledge for inference.

14. Small-Sample Corrections (e.g., Welch's t-test): Adjusts tests for small sample sizes.

15. Non-Parametric Tests (e.g., Median Test): Doesn't assume normality.

17.11 SUMMARY

In statistics, a large sample refers to a sample size that is sufficiently big to accurately represent the population from which it was drawn. The exact definition of "large" can vary depending on the context and the statistical analysis being performed, but here are some general guidelines:

1. Sample size greater than 30: In many statistical tests, a sample size of 30 or more is considered large enough to assume normality and apply parametric tests.

2. Sample size greater than 100: For more precise estimates and narrower confidence intervals, a sample size of 100 or more is often recommended.

3. Sample size at least 10% of the population: If the population size is known, a large sample is often defined as at least 10% of the population.

Small Sample refers to a statistical sample with a limited number of observations, typically fewer than 30. When dealing with small samples:

1. Central Limit Theorem (CLT) may not apply.

2. Distributional assumptions are crucial.

3. Statistical tests may have reduced power.

4. Confidence intervals may be wider.

17.12 TEST YOUR KNOWLEDGE

What is large sample?

What are the benefits of large sample?

Explain the limitation of large sample

Mention any two large sample test and explain the same with example

What is Small sample?

What are the benefits of Small sample?

Explain the limitation of Small sample

Mention any two Small sample test and explain the same with example

Highlight the different types of small sample

17.13 Further Readings

1. Sankalp Gaurav, Business Statistics, Agra book International , Agra
2. Sinha, V.C., Principles of Statistics.
3. Gupta, S.B., Principles of Statistics.
4. Monga, G.S., Elementary Statistics.

UNIT 18 NON PARAMETRIC TEST

UNIT STRUCTURE

OBJECTIVES

After going through this unit you should be able to know about the–

1. Concept of non parametric test
2. Different types of non parametric test and its application
3. How to calculate different types of non parametric test

18.1 NON PARAMETRIC TEST

A non-parametric test is a statistical test that doesn't require a specific distribution (like normality) of the data or population. These tests are often used when:

1. The data isn't normally distributed.

2. The sample size is small.

3. The data has outliers or extreme values.

4. The data is ordinal or categorical.

Non-parametric tests focus on the ranks or distributions of the data rather than the actual values. They're often used for hypothesis testing, such as comparing groups or assessing relationships. Examples of non-parametric tests include:

- Wilcoxon rank-sum test (Mann-Whitney U test)

- Wilcoxon signed-rank test

- Kruskal-Wallis H test

- Spearman's rank correlation coefficient

- Chi-squared test (for categorical data)

Non-parametric tests are useful because they're more flexible and robust than parametric tests, which assume specific distributions. However, they might be less powerful or efficient than parametric tests when the assumptions are met.

Advantages of Non Parametric Test:

1. No distributional assumptions: Non-parametric tests don't require normality or any specific distribution.

2. Robustness: They're less affected by outliers or extreme values.

3. Flexibility: Can be used with ordinal or categorical data.

4. Easy to understand: Results are often based on ranks or medians, making them more intuitive.

Disadvantages of Non Parametric Test:

1. Less powerful: Non-parametric tests might be less efficient than parametric tests when the assumptions are met.

2. Limited inference: Results might not provide detailed information about the population parameters.

3. Sample size limitations: Some non-parametric tests require larger sample sizes to be reliable.

Common applications of Non Parametric Test:

1. Comparing groups: Non-parametric tests are useful when comparing two or more groups with non-normal data.

2. Correlation analysis: Spearman's rank correlation coefficient is a popular non-parametric measure of correlation.

3. Ordinal data analysis: Non-parametric tests are suitable for ordinal data, like Likert scales or rankings.

Choosing a non-parametric test:

1. Consider the research question: What are you trying to test or compare?

2. Look at the data: Check for normality, outliers, and the type of data (ordinal, categorical, etc.).

3. Select the appropriate test: Based on the research question and data characteristics.

18.2 WILCOXON RANK-SUM TEST (MANN-WHITNEY U TEST)

The Wilcoxon rank-sum test (Mann-Whitney U test) is a non-parametric test used to compare the distributions of two independent groups or samples. It's a powerful alternative to the t-test when the data doesn't meet the assumptions of normality or equal variances.

Here are some key aspects of the Wilcoxon rank-sum test:

Null Hypothesis (H0): The two groups come from the same distribution.

Alternative Hypothesis (H1): The two groups come from different distributions.

Test Procedure:

1. Combine the data from both groups and rank them from smallest to largest.

2. Calculate the sum of ranks for each group.

3. Calculate the U statistic, which is the smaller of the two sums of ranks.

Interpretation:

- If the p-value is below the significance level (usually 0.05), reject H0 and conclude that the groups have different distributions.

- If the p-value is above the significance level, fail to reject H0 and conclude that the groups come from the same distribution.

Assumptions:

- Independent observations

- Identical shapes of the distributions (not necessarily normal)

- No ties or a small number of ties (if there are ties, use a tie-corrected version of the test)

Advantages:

- No assumption of normality

- Robust to outliers

- Can handle ordinal data

- Non-parametric, making it a good alternative to the t-test when assumptions are not met

Disadvantages:

- Less powerful than the t-test if the data is normally distributed

- May not be suitable for very small sample sizes

Software and Calculation:

- Most statistical software packages, such as R, Python, and SPSS, can perform the Wilcoxon rank-sum test.

- You can also calculate the test statistic and p-value manually using formulas.

18.3 THE WILCOXON SIGNED-RANK TEST

The Wilcoxon signed-rank test is a non-parametric test used to compare two related samples or repeated measurements. It's a powerful alternative to the paired t-test when the data doesn't meet the assumptions of normality or equal variances.

Here are some key aspects of the Wilcoxon signed-rank test:

Null Hypothesis (H0): The median difference between the two related samples is zero.

Alternative Hypothesis (H1): The median difference between the two related samples is not zero.

Test Procedure:

1. Calculate the differences between the paired observations.

2. Rank the absolute differences from smallest to largest.

3. Assign a positive sign to the rank if the difference is positive, and a negative sign if the difference is negative.

4. Calculate the sum of the positive ranks (W+).

5. Calculate the sum of the negative ranks (W-).

Test Statistic:

- The smaller of W+ or W- is the test statistic.

Interpretation:

- If the p-value is below the significance level (usually 0.05), reject H0 and conclude that the median difference is not zero.

- If the p-value is above the significance level, fail to reject H0 and conclude that the median difference is zero.

Assumptions:

- Paired observations

- Symmetrical distribution of differences (not necessarily normal)

- No ties or a small number of ties (if there are ties, use a tie-corrected version of the test)

Advantages:

- No assumption of normality

- Robust to outliers

- Can handle ordinal data

- Non-parametric, making it a good alternative to the paired t-test when assumptions are not met

Disadvantages:

- Less powerful than the paired t-test if the data is normally distributed

- May not be suitable for very small sample sizes

Here's a step-by-step guide to solving the Wilcoxon signed-rank test:

Step 1: Calculate the differences

Calculate the differences between the paired observations (d = x2 - x1).

Step 2: Rank the absolute differences

Rank the absolute differences from smallest to largest, ignoring the signs.

Step 3: Assign signs

Assign a positive sign to the rank if the difference is positive, and a negative sign if the difference is negative.

Step 4: Calculate W+ and W-

Calculate the sum of the positive ranks (W+) and the sum of the negative ranks (W-).

Step 5: Determine the test statistic

The smaller of W+ or W- is the test statistic (W).

Step 6: Calculate the p-value

Use a statistical table or software to find the p-value associated with the test statistic (W).

Step 7: Interpret the results

If the p-value is below the significance level (usually 0.05), reject the null hypothesis and conclude that the median difference is not zero.

Example:

X1	X2	d = X2-X1	Rank X	Sign
10	12	2	1	+
15	18	3	2.5	+
20	15	-5	4	-
25	28	3	2.5	+
30	20	-10	5	-

W+ = 1 + 2.5 + 2.5 = 6

W- = 4 + 5 = 9

W = min(6- 9) = -3

Using a statistical table or software, we find that the p-value associated with W = 5 is 0.031. Since $p < 0.05$, we reject the null hypothesis and conclude that the median difference is not zero.

18.4 THE KRUSKAL-WALLIS H TEST

The Kruskal-Wallis H test is a non-parametric statistical test used to compare the distributions of three or more independent groups. It's an extension of the Wilcoxon rank-sum test for more than two groups.

Here's a brief overview:

Null Hypothesis (H0): All groups come from the same distribution.

Alternative Hypothesis (H1): At least one group comes from a different distribution.

Test Procedure:

1. Rank all data from smallest to largest, ignoring group labels.

2. Calculate the sum of ranks for each group.

3. Calculate the test statistic (H) using the sum of ranks and group sizes.

Test Statistic (H):

$H = (12 / (N * (N + 1))) * (\sum(Ri^2 / ni)) - 3 * (N + 1)$

where:

- N = total sample size

- Ri = sum of ranks for group i

- ni = sample size of group i

Interpretation:

- If the p-value is below the significance level (usually 0.05), reject H0 and conclude that at least one group has a different distribution.

- If the p-value is above the significance level, fail to reject H0 and conclude that all groups come from the same distribution.

Assumptions:

- Independent observations

- Identical shapes of the distributions (not necessarily normal)

- No ties or a small number of ties (if there are ties, use a tie-corrected version of the test)

Advantages of Kruskal-Wallis H test

- Non-parametric, making it a good alternative to ANOVA when assumptions are not met

- Can handle ordinal data

- Robust to outliers

Disadvantages of Kruskal-Wallis H test:

- Less powerful than ANOVA if the data is normally distributed

- May not be suitable for very small sample sizes

Here's a step-by-step guide to solving the Kruskal-Wallis H test:

Step 1: Rank all data

Rank all data from smallest to largest, ignoring group labels.

Step 2: Calculate sum of ranks for each group

Calculate the sum of ranks for each group (Ri).

Step 3: Calculate group sizes

Calculate the sample size of each group (ni).

Step 4: Calculate the test statistic (H)

Calculate the test statistic (H) using the formula:

$H = (12 / (N * (N + 1))) * (\sum(Ri^2 / ni)) - 3 * (N + 1)$

where:

- N = total sample size

- Ri = sum of ranks for group i

- ni = sample size of group i

Step 5: Determine degrees of freedom

Determine the degrees of freedom (k - 1), where k is the number of groups.

Step 6: Calculate the p-value

Use a statistical table or software to find the p-value associated with the test statistic (H) and degrees of freedom.

Step 7: Interpret the results

If the p-value is below the significance level (usually 0.05), reject the null hypothesis and conclude that at least one group has a different distribution.

Example:

GROUP	DATA	RANK
1	10,12,15	1,2,3
2	18,20,22	4,5,6
3	25,28,30	,7,8,9

R1 = 1 + 2 + 3 = 6

R2 = 4+5+6 = 15

R3 -7+8+9 = 24

n1 = n2 = n3 = 3

N = 3 + 3 + 3 = 9

H = (12 / (9 * (9 + 1))) * ((6^2 / 3) + (15^2 / 3) + (24^2 / 3)) - 3 * (9 + 1)

H ≈ 6.33

Using a statistical table or software, we find that the p-value associated with H = 6.33 and k - 1 = 2 is 0.042. Since $p < 0.05$, we reject the null hypothesis and conclude that at least one group has a different distribution.

18.5 SPEARMAN'S RANK CORRELATION

Spearman's rank correlation coefficient (ρ) is a non-parametric measure of correlation between two variables. It assesses the strength and direction of the relationship between the rankings of two variables.

Here's a brief overview:

Formula:

ρ = 1 - (6 * Σ(d^2)) / (n * (n^2 - 1))

where:

- d = difference between ranks

- n = number of observations

Interpretation:

- ρ = 1: Perfect positive correlation

- ρ = -1: Perfect negative correlation

- ρ = 0: No correlation

Advantages of Spearman's rank correlation:

- Non-parametric, making it suitable for ordinal data or non-normal distributions

- Robust to outliers

Disadvantages of Spearman's rank correlation:

- Less powerful than Pearson's correlation coefficient for normally distributed data

- May not detect non-linear relationships

Example:

X	Y	RANK X	RANK Y	d
10	20	1	2	-1
15	25	2	4	-2
20	18	3	1	2

25	22	4	3	1
30	28	5	5	0
				0

$\rho = 1 - (6 * 0) / (5 * (5^2 - 1)) = 1$

In this example, there is a perfect positive correlation between X and Y.

18.6 CHI-SQUARE TEST

The Chi-square test is a statistical test used to determine whether there is a significant association between two categorical variables. It's commonly used to:

1. Test independence: Determine if two variables are independent or related.

2. Test goodness of fit: Determine if observed frequencies match expected frequencies.

Types of Chi-square tests:

1. Pearson's Chi-square test

2. Yates' corrected Chi-square test (for 2x2 tables)

3. Fisher's exact test (for small sample sizes)

Assumptions of Chi-square tests:

1. Independence: Observations are independent.

2. Random sampling: Sample is randomly selected.

3. Large enough sample size: Expected frequencies > 5.

Formula of Chi-square tests:

$\chi^2 = \Sigma [(\text{observed frequency} - \text{expected frequency})^2 / \text{expected frequency}]$

where:

- χ^2 = Chi-square statistic

- observed frequency = actual count

- expected frequency = count expected under independence

Interpretation:

- χ^2 value: Measures the difference between observed and expected frequencies.

- p-value: Probability of observing the χ^2 value (or more extreme) under independence.

- Degrees of freedom (df): Number of categories - 1.

Decision:

- If p-value ≤ α (significance level, usually 0.05), reject the null hypothesis (independence).

- If p-value > α, fail to reject the null hypothesis (independence).

Example of Chi-square tests:

	YES	NO	TOTAL
MALE	20	30	50
FEMALE	30	20	50
TOTAL	50	50	100

$\chi^2 = 10$, df = 1, p-value = 0.0016

Since p-value ≤ 0.05, reject the null hypothesis. There is a significant association between gender and the response variable.

18.7 SUMMARY

A non-parametric test is a statistical test that doesn't require a specific distribution (like normality) of the data or population.

The Wilcoxon rank-sum test (Mann-Whitney U test) is a non-parametric test used to compare the distributions of two independent groups or samples. It's a powerful alternative to the t-test when the data doesn't meet the assumptions of normality or equal variances.

The Wilcoxon signed-rank test is a non-parametric test used to compare two related samples or repeated measurements. It's a powerful alternative to the paired t-test when the data doesn't meet the assumptions of normality or equal variances.

The Kruskal-Wallis H test is a non-parametric statistical test used to compare the distributions of three or more independent groups. It's an extension of the Wilcoxon rank-sum test for more than two groups.

Spearman's rank correlation coefficient (ρ) is a non-parametric measure of correlation between two variables. It assesses the strength and direction of the relationship between the rankings of two variables. The Chi-square test is a statistical test used to determine whether there is a significant association between two categorical variables.

18.8 TEST YOUR KNOWLEDGE

1. What is non parametric test?

   .....
   .....
   .....

2. Explain wilcoxon rank sum test?

   .....
   .....
   .....

3. Explain the kruskal wallis H Test

   .....
   .....
   .....
   .....

4. With imaginary figures demonstrate spearman rank test

.....
.....
.....
.....

18.9 FURTHER READINGS

1. Sankalp Gaurav, Business Statistics, Agra Book International
2. Alhance D.N. and Alhance Beena, Fundamental of Statistics.
3. Monga, G.S., Elementary Statistics.
4. Gupta, S.B., Principles of Statistics.

UNIT – 19 CORRELATION AND REGRESSION

Unit Structure :

OBJECTIVES

After going through this unit you should be able to know about the–

1. Karl Pearson's Coefficient of Correlation.
2. Spearman's Rank Correlation.
3. Regression.

19.1. INTRODUCTION

Meaning : In various types of analyses, we have confined ourselves to such series where various Items assumed different values of one variable. We have discussed how measures of central tendency and measures of dispersion and skewness are calculated in such cases for purposes of comparison and analysis. With the help of these measures, such data can be easily understood. There can, however, be such series also where each item assumes the values of two or more variables. For example, if the heights and weights of a group of persons are measured, we shall get such series where each member of the group would assume two values—one relating to height and the other relating to weight. If, besides heights and weights, the chest measurements were also taken, each member of the group would assume three values relating to three different variables. In such cases, we can calculate averages, dispersion and skewness, etc., in accordance with the rules given in the previous chapters.

But, sometimes, it appears that the values of the various variables so obtained are interrelated. It is likely that such relationship may be obtained in two series relating to the heights and weights of a group of persons. It may be observed that weights increase with increase in

heights so that tall people are heavier than short sized people. Similarly, if the data are collected about the prices of a commodity and the quantities sold at different prices, two series would be obtained. One variable would be the various prices of the commodity, and the other variable would be the quantities sold at these prices. In two such series, we are again likely to find some relationship. With increase in the price of the commodity, the quantity sold is bound to decrease. We can, thus, conclude that there is some relationship between price and demand. Such relationships can be found in many types of series, for example, prices and supply, heights and weights of persons, prices of sugar and sugarcane, ages of husbands and wives, etc.

The term correlation (or co-variation) indicates the relationship between two such variables in which, with changes in the values of one variable, the values of the other variable also change if the two variables move together.

19.2. DEFINITION

Some important definitions of correlation are given below:

(1) "If two or more quantities vary in sympathy so that movements in the one tend to be accompanied by corresponding movements in the other(s), then they are said to be correlated". — *L.R. Connor*

(2) "Correlation means that between two series or groups of data, there exists some casual connection" . — *W.L. King*

(3) "Correlation analysis attempts to determine the 'degree of relationship' between variables. — *Ya Lun Chow*

(4) "When the relationship of a quantitative nature, the appropriate statistical tool for discovering and measuring the relationship and expressing it in a brief formula is known as correlation." — *Croxton & Cowden*

The above definitions make it clear that the tern correlation refers to the study of relationship between two or more variables.

19.3. TYPES OF CORRELATION

Correlation can be :

(i) Positive or Negative;

(ii) Simple, Multiple or Partial;

(iii) Linear or Non-linear.

(i) Positive or Negative Correlation

Correlation can be either positive or negative. When the values of two variables move in the same direction, i.e., when an increase in the value of one variable is associated with an increase in the value of the other variable and a decrease in the value of one variable is associated with the decrease in the value of the other variable, correlation is to be positive.

If, on the other hand, the values of two variables move in opposite directions, so that with an increase in the values of one variable, the values of the other variable decrease, and with a decrease in the values of one variable the values of the other variable increase, correlation is said to be

negative. There are some data in which correlation is generally positive while, in others, it is negative. Thus, generally, price and supply are positively correlated. When prices go up, supply also increases and with the fall in prices, supply also decreases. The correlation between price and demand is generally negative. With an increase in price, the demand goes down and with a decrease in price the demand generally goes up. Demand curve is downward sloping, whereas, supply curve is upward sloping.

(ii) Simple, Multiple and Partial Correlation

In simple correlation, we study only two variables say, price and demand. In multiple correlation, we study together the relationship between three or more factors like production, rainfall and use of fertilizers. In partial correlation, though more than two factors are involved but correlation is studied only between two factors and the other factors are assumed to be constant.

(iii) Linear and Non-linear (Curvi-linear) Correlation

The correlation between two variables is said to be linear if corresponding to a unit change in the value of one variable, there is a constant change in the value of the other variable, i.e., in case of linear correlation the relation between the variables x and y is of the type

$$y = a + bx.$$

If $a = 0$, the relation becomes $y = bx$ in such cases, the values of the variables are in constant ratio. The correlation between two variables is said to be non-linear or curvilinear if corresponding to a unit change in the value of one variable the other variable does not change at a constant rate but at a fluctuating rate.

19.4. METHODS OF DETERMINING CORRELATION

The various methods by which correlation studies are made, are as follows:

(i) Scatter Diagram

(ii) Correlation Graph

(iii) Coefficient of Correlation

(iv) Coefficient of Correlation by Rank Differences

(v) Coefficient of Concurrent Deviation

(vi) Method of Least Square.

1. Coefficient of Correlation

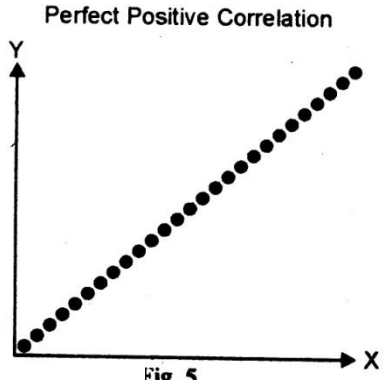
Coefficient of correlation is calculated to study the extent or degree of correlation between two variables. As has been said, earlier, the fact that there is correlation between two variables does not mean that their relationship is functional or constant. If the value of a variable is known, it is not always possible to obtain the exact value of the other variable. This can be done only where there is linear relationship between the two variables. There are a few series in which linear relationship exists, e.g., natural numbers and their squares or square roots would always give a linear relationship. Similarly, linear relationship would be obtained between two series, one relating to radii of various. circles and the other relating to their areas. In economic data, such relationships are rarely found. No doubt, demand would fall with an increase in price, but the relationship is not functional. There is no constant ratio between the variation of the two series rating to price and

demand.

Perfect Correlation: If the relationship between two variables is such that with an increase in the value of one, value of the other increases or decreases, in a fixed proportion, correlation between them is said to be perfect. If both the series move in the same direction and the variations are proportionate, there would be perfect positive correlation between them. If, on the other hand, the two series move in reverse directions, and the variations in their values are always proportionate, it is an example of perfect negative correlation. It is also likely that there may be no relationship between the variations of the two series in which case there is said to be no correlation between them.

As has been said earlier, in economic data, perfect positive or negative correlation is usually not found, as the relationship between economic series is rarely functional. In such data, correlation is not perfect as the related series are not completely dependent on each other. Perfect correlation is obtained when there is complete mutual dependence between the two series.

It would be observed figs. 1 and 2 that all corresponding values of x and y are in a straight line. figure 1 indicates perfect positive correlation between x and y as the variation in the values of the two series are always in a fixed proportion and they move in the same direction. figure 2, on the other hand, shows a perfect negative correlation between x and y as the variations between their values are in a constant ratio and two series move in reverse directions.



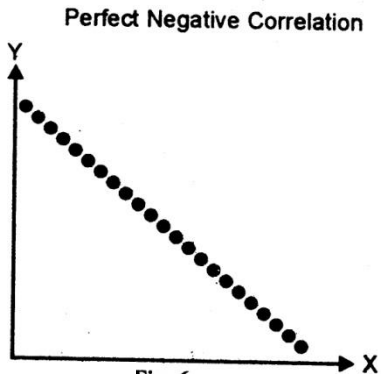


Fig. 1 Fig. 2

After knowing this, it is necessary to obtain such a measure of correlation which can accurately indicate the degree of correlation in quantitative terms. The measure should be such that its extreme values represent perfect positive and perfect negative correlations and the value in the middle, absence of correlation. Such a measure is given by the coefficient of correlation.

The coefficient of correlation which we are going to discuss in the following pages always varies between the two limits of + 1 and – 1. When there is perfect positive correlation, its value is + 1 and when there is perfect negative correlation its value is – 1. Its midpoint is 0, which indicates absence of correlation. As the value of this coefficient decreases from the upper limit of + 1, the extent of positive correlation between the two variables also declines. When it reaches the value of 0, it indicates complete absence of correlation and when it goes further down in negative values (less than zero) it indicates negative correlation. When it reaches the other limit of – 1 there is evidence of perfect negative correlation between the two series.

The above mentioned points can be studied from the graphs which have been given so far. When the values of the variable are like those given in figure 1, there is perfect positive correlation or the value of the coefficient of correlation is + 1, when they are like those given in figure 1, there is positive correlation but it is not perfect, or the value of the coefficient of correlation is less than + 1 but more

than 0. When the values are like those given in Figure 3, there is no correlation between the data or the value of the coefficient of correlation is 0; when the values are like those given in Figure 2, there is negative correlation though not perfect, which means that the value of the coefficient of correlation would be more than 0 (on the negative side) but less than – 1. If, however, the values of the variable are like those given in figure 2, there. is perfect negative correlation or in other words, the value of the coefficient of correlation would be – 1.

19.5. CALCULATION OF COEFFICIENT OF CORRELATION

(Karl Pearson's Formula)

Karl Pearson, the great biologist and statistician, has given a formula for the calculation of coefficient of correlation. According to it the coefficient of correlation of two variables is obtained by dividing the sum of the products of the corresponding deviations of the various items of the two series from their respective means by the product of their standard deviations and the number of pairs of observations.

If X1, X2, X3, ... Xn are the values of the first variable and Y1, Y2, Y3, Yn are the values of the second variable, then:

Correlation coefficient (r) = $\dfrac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n\,\sigma_x\sigma_y}$ or $r = \dfrac{\sum dx\,dy}{n\,\sigma_x\sigma_y}$

Thus, if x1, x2, x3, ... xn are the deviations of various items of the first variable from their mean value and y1, y2, y3, ... yn are the corresponding deviations of the second variable from its mean value, the sum of the products of these corresponding deviations would be ∑xy. If, further, the standard deviations of the two variables are respectively □x, □y and if n is the number of pairs of observations, Karl Pearson's coefficient of correlation represented by r would be :

$$r = \frac{\sum_{i=1}^{n} x_i y_i}{n\,\sigma_x\sigma_y} \quad \text{.....(i)}$$

It is clear from the above formula that if ∑xy is positive, the coefficient of correlation would also be a positive figure indicating positive correlation between the two series. If, on the other hand, ∑xy is negative, coefficient of correlation would also be negative, indicating that the 'correlation between the two series is negative, ∑xy would be positive, if generally, positive and negative deviations in one series are associated with positive and negative deviations in the other series also, i.e., the deviations in both the series have the same sign. The value of ∑xy would be negative, if the positive deviations of one variable are associated with the negative deviations in the other variable and vice versa. The deviations is both the series are of opposite sign. If positive and negative deviations of one variable are indifferently associated with the deviations of the other variable, the value of would be 0 or near it, indicating absence of correlation between the two series. The value of this coefficient of correlation always lies between + 1 and – 1. It cannot exceed unity.

The above formula of Karl Pearson is based on the study of covariance between two series. The covariance between two series is written as follows :

$$\text{Covariance (X, Y)} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n}$$

To study the correlation, the covariance of the two series is divided by the product of their standard deviations. Thus,

$$r = \frac{\text{covariance of the two series}}{(\text{variance of series1})(\text{variance of series2})} = \frac{\text{Covariance X, Y}}{\sigma_x \times \sigma_y}$$

Since the S.D. is independent of the change of origin $\Box x = \Box x$, $\Box y = \Box y$

$$\frac{\sum xy}{n\,\sigma_x \sigma_y} \quad \text{Where } x = X - \bar{X},\ y = Y - \bar{Y}$$

This formula is known as the Product moment formula of Coefficient of Correlation.

Calculation of the Pearson's Coefficient of Correlation

$$r = \frac{\sum(X-\bar{X})(Y-\bar{Y})}{\sqrt[n]{\frac{\sum(X-\bar{X})^2}{n}}\sqrt{\frac{\sum(Y-\bar{Y})^2}{n}}} = \frac{\sum(X-\bar{X})(Y-\bar{Y})}{\sqrt{\frac{\sum X^2}{n} - \left(\frac{\sum X}{n}\right)^2}\sqrt{\frac{\sum Y^2}{n} - \left(\frac{\sum Y}{n}\right)^2}}$$

Now, $$\frac{\sum(X-\bar{X})(Y-\bar{Y})}{n} = \frac{\sum XY - \sum X\bar{Y} - \sum \bar{X}Y + \sum \bar{X}\bar{Y}}{n}$$

$$= \frac{\sum XY}{n} - \bar{Y}\frac{\sum X}{n} - \bar{X}\frac{\sum Y}{n} + \frac{\sum \bar{X}\bar{Y}}{n}$$

$$= \frac{\sum XY}{n} - \bar{X}\bar{Y} - \bar{X}\bar{Y} + \bar{X}\bar{Y} = \frac{\sum XY}{n} - \left(\frac{\sum X}{n}\right)\left(\frac{\sum Y}{n}\right)$$

$\because$ $$\bar{X} = \frac{\sum X}{n}, \quad \bar{Y} = \frac{\sum Y}{n}$$

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

If the deviations of x-series are taken from any value A and that of Y-series taken from the value B, Then.

$$X - \bar{X} = X - \bar{X} + A - A = (X - A) - (\bar{X} - A) = dx - \bar{d}x$$

$$Y - \bar{X} = Y - \bar{Y} + B - B = (Y - B) - (\bar{Y} - B) = dy - \bar{d}y$$

By making the x substitution in A we get

$$r = \frac{\sum\left(dx - \bar{d}x\right)\sum\left(dy - \bar{d}y\right)}{n\sqrt{\frac{\sum\left(dx - \bar{d}x\right)^2}{n}}\sqrt{\frac{\sum\left(dy - \bar{d}y\right)^2}{n}}}$$

$$r = \frac{\frac{\sum d_x d_y}{n} - \left(\frac{\sum d_x}{n}\right)\left(\frac{\sum d_y}{n}\right)}{n\sqrt{\frac{\sum d_x^2}{n} - \left(\frac{\sum d_x}{n}\right)^2}\sqrt{\frac{\sum d_y^2}{n} - \left(\frac{\sum d_y}{n}\right)^2}}$$

If the deviations are taken from their respective actual means, then

$$r = \frac{\sum xy}{n\sqrt{\sum x^2}\sqrt{\sum y^2}}$$

where $x = X - \overline{X}$ $y = Y - \overline{Y}$

Correlation Coefficient is independent of the change of origin as well as change of scale.

The following solved examples would illustrate the use of the above rules.

Example. Find out the coefficient of correlation between the sales and expenses of the following 10 firms (figure in '000 Rs.)

Firms :	1	2	3	4	5	6	7	8	9	10
Sales :	50	50	55	60	65	65	65	60	60	50
Expenses :	11	13	14	16	16	15	15	14	13	13

Solution

Calculation of Coefficient of Correlation

Sales (X)	**Dev. from mean (58) (x)**	**Dev. square (x^2)**	**Expen-ses (Y)**	**Dev. from mean (14) (y)**	**Dev. square (y^2)**	**Product of deviation (xy)**
50	– 8	64	11	– 3	9	+ 24
50	– 8	64	13	– 1	1	+ 8
55	– 3	9	14	0	+ 0	0
60	+ 2	4	16	+ 2	4	+ 4

65	+ 7	49	16	+ 2	4	+ 14
65	+ 7	49	15	+ 1	1	+ 7
65	+ 7	49	15	+ 1	1	+ 7
60	+ 2	4	14	0	0	0
60	+ 2	4	13	– 1	1	– 2
50	– 8	64	13	– 1	1	+ 8
$\sum X = 580$ $n = 10$	$\sum x = 0$	$\sum x^2 = 360$	$\sum Y = 140$	$\sum y = 0$	$\sum y^2 = 22$	$\sum xy = 70$

$$\text{Mean of X} = \frac{\sum X}{n} = \frac{580}{10} = 58$$

$$\text{Mean of Y} = \frac{\sum Y}{n} = \frac{140}{10} = 14$$

$$r = \frac{\sum xy}{n\sqrt{\frac{\sum x^2}{n}}\sqrt{\frac{\sum y^2}{n}}} = \frac{+70}{10\sqrt{\frac{360}{10}}\sqrt{\frac{22}{10}}}$$

$$= \frac{+70}{10\sqrt{36}\sqrt{2.2}} = 0.786$$

Alternatively : $$r = \frac{\sum xy}{\sqrt{\sum x^2 \times \sum y^2}} = \frac{+70}{\sqrt{360 \times 22}} = +0.786$$

Example. Calculate Coefficient of Correlation between the values of X and Y given below :

Values of X:	65	66	67	67	68	69	70	72
Values of Y:	67	68	65	68	72	72	69	71

Solution

Calculation of Coefficient of Correlation

(X)	**(Y)**	**(X^2)**	**(Y^2)**	**(XY)**
65	67	4225	4489	4355
66	68	4356	4624	4488
67	65	4489	4225	4355

67	68	4489	4624	4556
68	72	4624	5184	4896
69	72	4761	5184	4968
70	69	4900	4761	4830
72	71	5184	5041	5112
$\sum(X) = 544$	$\sum(Y) = 552$	$\sum(X)^2 = 37028$	$\sum(Y)^2 = 38132$	$\sum(XY) = 37560$

Coefficient of correlation :

$$= \frac{\frac{\sum XY}{N} - \left(\frac{\sum X}{N}\right)\left(\frac{\sum Y}{N}\right)}{\sqrt{\sum \frac{X^2}{N} - \left(\frac{\sum X}{N}\right)^2}\sqrt{\sum \frac{Y^2}{N} - \left(\frac{\sum Y}{N}\right)^2}}$$

$$= \frac{\frac{37560}{8} - \frac{544}{8} \times \frac{552}{8}}{\sqrt{\frac{37028}{8} - \left(\frac{544}{8}\right)^2}\sqrt{\frac{38132}{8} - \left(\frac{552}{8}\right)^2}}$$

$$\frac{4695 - 4692}{\sqrt{4628.5 - 4624}\sqrt{4766.5 - 4761}} = \frac{3}{4.975} = 0.6030$$

Assumption of Pearsonian Correlation

The Pearsonian coefficient of correlation rests on two assumptions.

The first is that a large number of independent contributory causes are operating in each of the two series correlated so as to produce normal or probability distribution. We know that such causes always operate in chance phenomena like tossing of coin or throw of a dice. They also operate in other types of data. For example, such forces are usually found operating in phenomena like indices of price and supply, ages of husbands and wives and heights of fathers and sons, etc.

The second assumption is that the forces so operating are not independent of each other but are related in a casual fashion. If the forces are entirely independent and unrelated, there cannot be any correlation between the two series. The forces must be common to both the series. The height of an individual during the last ten years may show an upward trend and his income during this period may also show a similar tendency but there cannot be any correlation between the two series because the forces affecting the two series are entirely unconnected with each other. If the coefficient of correlation in such series is calculated, it may even be +.8 indicating a very high degree of positive correlation, but such correlation is usually termed nonsense correlation because the two series are affected by such sets of forces which are entirely unconnected with each other.

In the words of Karl Pearson, "The sizes of the complex of organs (something measurable) are determined by a great variety of independent contributing causes, for example, climate, nourishment, physical training and innumerable other causes which cannot be individually observed or their

effects measured." Karl Pearson further observes, "The variations in intensity of the contributory causes are small as compared with their absolute intensity and these variations follow the normal law of distribution."

19.6. RANK CORRELATION

Sometimes, such problems are faced where it is possible to arrange the various items of a series in serial order but the quantitative measurement of their value is difficult; for example, it is possible for a class teacher to arrange his students in ascending or descending order of intelligence, even though intelligence cannot be measured quantitatively. No doubt, the quantitative study about the intelligence of students can be made by holding an examination and assigning them marks, but this method can never be said to be infallible. There are many such attributes which are incapable of quantitative measurements; for example, honesty, character, morality, beauty etc.

If it is desired to have a study of association between two such attributes, say, intelligence and beauty, the Karl Pearson's Coefficient of Correlation cannot be calculated as these attributes cannot be assigned definite values. However, there is a method by which we can study correlation between such attributes. This method was developed by the British psychologist, Charles Edward Spearman, in the year 1904.

In this method, x and y variables denote the rank of the attributes A and B. In case of a study of correlation between intelligence and beauty, we can pick up 10 or 20 or any other number of individuals and first arrange them in order of their rank according to intelligence-beginning with the most intelligent person whose rank would be 1 and going down in order till the rank of the last person is indicated. Similarly, we can arrange these individuals again in order of rank according to beauty the most beautiful person getting Rank No.1 and going down to the last person who is the least beautiful.

In this way, we will have two sets of ranks for these two attributes. If 10 individuals are arranged like this, we will have ranks from 1 to 10 for each attribute.

Once this is done, we can find out a Coefficient of Correlation between these two series by the Spearman's Rank Correlation formula which is as under :

$$r_s = 1 - \frac{6\sum d^2}{n\left(n^2 - 1\right)} \quad \text{or} \quad 1 - \frac{6\sum d^2}{n^3 - n} \quad \text{or} \quad r = 1 - \frac{6\sum d^2}{N\left(N^2 - 1\right)}$$

where r_s denotes the Spearman's Rank Correlation and d denotes the difference of the corresponding ranks of the same individual in the two attributes, and n denotes the number of pairs.

The value of the rank correlation coefficient is interpreted in the same way as that of the Pearson's Coefficient of Correlation. Its value also ranges between + 1 and – 1. When r_s is + 1 it indicates complete agreement in the order of ranks between the two attributes (the most intelligent being the most beautiful also, and so on). If correlation is – 1 it indicates complete disagreement in the order of ranks (the most intelligent being the one who is the least beautiful).

The rank correlation coefficient will be equal to – 1 if the ranks assigned by the judges are exactly inverse i.e., an individual who gets the highest score from one judge, gets the lowest scores from the other and the individual getting the lowest score from one judge gets the highest from the other and so on.

There are two types of problems in calculating this coefficient :

(A) When actual ranks are given.

(B) When actual ranks are not given.

In each of these two types of problem, a difficulty arises when ranks of two individuals are the same. Such problems need a modification in the formula given above. (See example as Equal ranks).

(A) Where Ranks are Given

Where actual ranks of the items are given the steps that have to be taken to get the coefficient are as follows :

(i) Compute the difference of ranks (R_1 — R_2) and denote them by d.

(ii) Compute d^2 and total them to get $\sum d^2$.

(iii) Use the formula given below to get the Coefficient:

$$r_s = 1 - \frac{6\sum d^2}{n^3 - n}$$

The following examples will illustrate the above method.

Example. The values of the same 15 students in two subjects A and B are given below; the two numbers within the brackets denoting the ranks of the same student in A and B, respectively:

(1, 10) (2,7) (3,2) (4, 6) (5, 4) (6,8) (7,3) (8, 1)

(9, 11) (10, 15) (11,9) (12,5) (13, 14) (14, 12) (15, 13)

Use Spearman's formula to find the rank Correlation Coefficient.

Solution.

Calculation of Rank Correlation Coefficient

Rank in A R_1	**Rank in B** R_2	$(R_1 - R_2)$ **d**	d^2
1	10	– 9	81
2	7	– 5	25
3	2	– 1	1
4	6	– 2	4
5	4	– 1	1
6	8	– 2	4
7	3	+ 4	16
8	1	– 7	49
9	11	– 2	4
10	15	– 5	25

11	9	– 2	4
12	5	– 7	49
13	14	– 1	1
14	12	– 2	4
15	13	– 2	4
n = 15		$\sum d = 0$	$\sum d^2 = 272$

Spearman's Coefficient of Correlation or

$$r_s = 1 - \frac{6\sum d^2}{n^2 - n}$$

Substituting the values, we get :

$$r_s = 1 - \frac{6 \times 272}{15^2 - 15} = 1 - \frac{1632}{3375 - 15} = 1 - \frac{1632}{3360} = 1 - \frac{17}{35} = \frac{18}{35} = +0.51$$

Example. Calculate the coefficient of rank correlation from the following data :

X	60	34	40	50	45	41	22	43	42	66	64	46
Y	75	32	34	40	45	33	12	30	36	72	41	57

Solution

Calculation of Coefficient of Rank Correlation.

X	**Rank 1**	**y**	**Rank 2 of Ranks (d)**	**Differ-ence**	**d^2**
60	3	75	1	+ 2	4
34	11	32	10	+ 1	1
40	10	34	8	+ 2	4
50	4	40	6	– 2	4
45	6	45	4	+ 2	4
41	9	33	9	0	0
22	12	12	12	0	0
43	7	30	11	– 4	16
42	8	36	7	+ 1	1

66 64 46	1 2 5	72 41 57	2 5 3	– 1 – 3 + 2	1 9 4
n = 12				0	$\sum d^2 = 48$

Coefficient of rank correlation or

$$r_s = 1 - \frac{6\sum d^2}{n\left(n^2 - 1\right)} = 1 - \frac{6(48)}{12\left(12^2 = 1\right)} = 1 - \frac{288}{1716} = \frac{1428}{1716} = 0.82$$

Equal Ranks

Sometimes, where there is more than one item with the same value, a common rank is given to such items. This rank is the average of the ranks which these items would have got had they differed slightly from each other. When this is d(,me, the coefficient of rank correlation needs some correction, because the above formula is based on the supposition that no item is repeated, the ranks of various items are different and that no rank is given to more than one item.

If in a series, there are 'm' items whose ranks are common, then for correction of the coefficient of rank correlation 1/12 ($m^3 - m$) is added to the value of ($\sum d^2$). If there are more than one such groups of items with common rank, this value is added as many times as the number of such groups.

The formula for the calculation of Rank Correlation is, thus, modified in the following manner:

$$r_s = 1 - \frac{6\left[\sum d^2 + \sum \frac{1}{12}\left(m^3 - m\right)\right]}{n^3 - n}$$

The following examples would illustrate the use of this modified formula :

Example. Calculate the coefficient of rank correlation from the following data:

X	48	33	40	9	16	16	65	24	46	57
Y	13	13	24	6	15	4	20	9	6	19

Solution.

Calculation of Coefficient of Rank Correlation

X	**Rank 1**	**Y**	**Rank 2**	**Diff. of Ranks (d)**	**d^2**

48	3	13	5.5	– 2.5	6.25
33	5	13	5.5	– 0.5	.25
40	4	24	1	– 3.0	9.00
9	10	6	8.5	– 1.5	2.25
16	8	15	4	+ 4.0	16.00
16	8	4	40	– 2.0	4.00
65	1	20	2	– 1.0	1.00
24	6	9	7	– 1.0	1.00
16	8	6	8.5	– 0.5	.25
57	2	19	3	– 1.0	1.00
n = 10		**n = 10**		□	**$\sum d^2$ = 41.00**

In the above table, in X-series, figure 16 occurs three times. The rank of all these items is 8 which is the average of 7, 8, and 9 — the ranks which these items would had there been some difference between their values. In Y-series figures 13 and 6 both occur two times. Their ranks are respectively 5.5 and 8.5. Due to these common ranks, the coefficient of rank correlation would have to be corrected.

For correction, we shall add [1/12 ($m^3 - m$)] to the value of [$\sum d^2$]. In X-series, this value would be equal to [1/12 ($3^3 - 3$)] as the value 16 has occurred three times in this series. In Y-series, there are two such groups of common ranks. In the first group, this correction would be [1/12 ($2^3 - 2$)] as the value 13 has occurred twice and for the second group also the correction value would be [1/12 ($2^3 - 2$)] as the value has also occurred twice in this series.

Spearman's Coefficient of Rank Correlation or

$$r_s = 1 - \frac{6\left[\sum d^2 + \frac{1}{12}\left(m_1^3 - m_1\right) + \frac{1}{12}\left(m_2^3 - m_2\right) + \ldots\right]}{n^3 - n}$$

$$= 1 - \frac{6\left[41 + \frac{1}{12}\left(3^3 - 3\right) + \frac{1}{12}\left(2^3 - 2\right) + \frac{1}{12}\left(2^3 - 2\right)\right]}{10^3 - 10}$$

$$= 1 - \frac{6\left(41 + 2 + 0.5 + 0.5\right)}{990} = 1 - \frac{264}{990} = \frac{726}{990} = +0.73$$

Example. A firm, not sure of the response to its product in ten different colour shades, decides to produce them in those colour shades.

The two judges rank the 10 colours in the following order :

Colour No.	1	2	3	4	5	6	7	8	9	10
Ranking by Judge I	6	4	3	1	2	7	9	8	10	5
Judge II	4	1	6	7	5	8	10	9	3	2

Is there any agreement between the two judges, to allow the introduction of the product by the firm in the market?

D	6–4	4–1	3–6	1–7	2–5	7–8	9–10	8–9	10–3	5–2
|D|	2	3	3	6	3	1	1	1	7	3
D^2	4	9	9	36	9	1	1	1	49	9
$\sum D^2 = 128$										

$$\text{Rank correlation} = 1-\frac{6\sum d^2}{N\left(N^2-1\right)} = 1-\frac{6\times 128}{10\left(100-1\right)} = 1-0.78 = 0.22$$ **Ans.**

Since the rank correlation is low, there is not much agreement between the two judges, and the firm may not produce the product to introduce it in the market.

Merits and Demerits of the Rank Correlation

Merits

(1) Since in this method $\sum D$ or the sum of the differences between R_1 and R_2 is always equal to zero, it provides a check on the calculation.

(2) Since Spearman's Rank Correlation is the same thing as Karl Pearson's Coefficient of Correlation between ranks, it can be interpreted in the same way as Karl Pearson's Coefficient of correlation.

(3) Rank correlation unlike Karl Pearson's Coefficient of Correlation does not assume normality in the universe from which the sample has been taken.

(4) Rank Correlation is very easy to understand and apply. However, Pearson's Coefficient is based on a set of full information while Spearman's Coefficient is based only on the ranks. The values of r obtained by these two methods would generally differ.

(5) Spearman's Rank method is the only way of studying correlation between qualitative data which cannot be measured in figures but can be arranged in serial order.

Demerits

(1) The method cannot be used in two-way frequency tables or bivariable frequency distribution.

(2) It can be conveniently used only when n is small, say, 30, otherwise calculation becomes tedious.

19.7. REGRESSION

Introduction

The dictionary meaning of the word regression is 'stepping back' or 'going back'. The use of this word dates back to the time of early studies made by **Francis Galton** in the latter half of the nineteenth century. He studied the relationship between the heights of fathers and their sons, and arrived at some very interesting conclusions. which are given below :

(i) Tall fathers have tall sons and short fathers have short sons.

(ii) The mean height of the sons of tall fathers is less than the mean height of their fathers.

(iii) The mean height of the sons of short fathers is more than the mean height of their fathers.

(iv) Galton found that the deviations in the mean height of the sons from the mean height of the race was less than the deviations in the mean height of the fathers from the mean height of the race. When the fathers move above the mean or below the mean, the sons tended to go back or regress towards the mean.

Regression, thus, implies going back or returning towards the mean. Galton studied the average relationship between these two variables graphically and called the line describing the relationship the line of regression.

Regression lines, thus, study the average relationship between two series and throw light on their Covariance. If the Coefficient of Correlation between the heights of fathers and sons is +0.7 it means that if a group of fathers have heights which are more than the average by x inches, their sons could have heights which would be more than average by 0.7 inches. Thus, the height of the sons regresses towards the mean. The study of this tendency is the subject matter of regression.

At this stage, we shall examine some definitions of this term.

1. "Regression is the measure of the average relationship between two or more variables in terms of the original units of the data."

2. "Regression analysis attempts to establish the, nature of the relationship' between variables. that is, to study the functional relationship between the variables and thereby provide a mechanism for predicting, or forecasting."

19.8. UTILITY OF REGRESSION

The above definitions make it clear that regression analysis is done for estimating or predicting the unknown value of one variable from the known value of the other variable. This is a very useful statistical tool which is used both in natural and social sciences.

In the field of business, this tool of statistical analysis is very widely used. Businessmen are interested in predicting future production, consumption, investment, prices, profits, sales, etc. In fact, the success of a businessman depends on the correctness of the various estimates that he is required to make. In sociological studies and in the field of economic planning, projections of population, birth rates, death rates and other similar variables are of great use.

In our day to day life, we come across many variables which are interrelated. For example, with a rise in price, the demand of a commodity goes down. or with better monsoons the output of agricultural product increase, or the effect of expenditure on publicity may lead to a rise in the volume of sales. With the help of regression analysis, we can estimate or predict the effect of one

variable on the other. e.g., we can predict the fall in demand when the price rises by a particular amount. However. in social sciences, there is multiple causation which means that a large number of factors affect various variables. The regression study which confines itself to a study of only two variables is called simple regression. The regression analysis which studies more than two variables at a time is called multiple regression.

In the simple regression analysis, there are two variables—one of which is known as an 'independent variable' or 'regresser' or predictor or explanator. On the basis of the values of this variable, the values of the other variable are predicted. This other variable whose values are predicted is called the 'dependent' or 'regressed' or explained variable.

With the help of regression studies, we can also calculate the value of the Coefficient of correlation. The Coefficient of determination [square of coefficient of correlation] which measures the effect of the independent variable on the dependent variable gives us an indication about the predictive value of the regression studies.

19.9. METHODS OF STUDYING REGRESSION

Broadly speaking, regression can be studied either :

(i) Graphically or

(ii) Algebraically

(i) Graphic Study of Regression

When regression is studied with the help of graphic methods, we have to draw a scatter diagram. A scatter diagram contains one point for one pair of values of X and Y variable. Usually, X variable is shown on the horizontal scale and Y variable on the vertical scale. When all related pairs of values have been plotted on a scatter diagram, we have to draw two regression lines to predict the values of X and Y variables. The regression line which is used to predict the values of Y for a value of X is called the Regression line of Y on X. Similarly, the regression line which is used to predict a value of X for a value of Y is called Regression line of X on Y. If the coefficient of correlation between X and Y is perfect, i.e., its value is either +1 or – 1, there will be only one regression line as the variations in the two series in such cases always increase or decrease by a constant figure. In other words, we say that the two regression lines will be identical if the correlation between the two variables is perfect.

If the graph between the values of a dependent variable and independent variable is a straight line, then the regression is called Linear Regression. If, however, the relationship between the two variables is not in the form of a straight tine but have some other functional relationship like $Y = X^2$, then regression between the variables is Non-linear regression. In this chapter, we are studying only linear regression.

How to draw linear regression lines

Regression lines can be drawn by :

(a) freehand curve method, or

(b) by the Method of Least Squares

Freehand Curve Method

In the freehand curve method, we first plot the pairs of the values of X and Y in the form of a scatter

diagram-one point for one pair of values. After this, we draw two free hand straight lines. One of these lines is drawn in such a way that the positive deviations of Y-series from its mean are cancelled by the negative deviations. The sum of the deviations on one side of the line is equal to the sum of deviations on the other side. This will be the regression line of Y on X. The other regression line would be drawn in such a way that the positive deviations of X-series from its mean would cancel the negative deviations. This regression line would be called the regression line of X on Y. The two regression lines would cut each other at the point the coordinates of which represent the means of two series. If there is perfect positive or negative correlations between the two variables, there will be only one regression line.

However, it is very difficult to draw regression lines by the freehand curve method. Usually, a piece of thread is repeatedly adjusted in such a manner in the scatter diagram that the positive and negative deviations cancel each other. Once these lines are drawn, we can predict or estimate the values of Y from the Regression line of Y on X and, similarly, the values of X can be predicted from the regression line X on Y.

Method of Least Squares

In order to avoid the difficulties associated with the drawing of regression lines by the freehand curve method. a mathematical relationship is established between the movements of X and Y series and algebraic equations are obtained to represents the relative movements of X and Y series.

In this method, we minimise the Sum of Squares of the deviations between the given values of a variable and its estimated values given by the line of the best fit. Line of Regression of Y on X is the line which gives the best estimate for the value of Y for a specified value of X and, similarly, the line of regression of X on Y is the line which gives the best estimate for the value of X for a specified value of Y.

If the values of Y are plotted on the Y axis (i.e., the vertical axis), then the regression line of Y on X will be such which minimises the sum of the squares of the vertical deviations. Similarly, if the values of X are plotted on the X axis (i.e., the horizontal axis) the regression line of X on Y will be such which minimises the sum of the squares of the horizontal deviations.

We have briefly discussed the method of least squares in the chapter on Correlation and pointed out that the line of the best fit is obtained by the equation of straight line $Y = a + bX$ and that in the method of least squares, this line is obtained with the help of the following two normal equations :

$$\sum Y = na + b (\sum X)$$

$$\sum XY = a \sum (X) + b \sum (X^2)$$

If the values of X and Y variables are substituted in the above equations, we get the values of a and b by solving these equations and thus, get the regression line of Y on X. Here, Y is the dependent variable and X the independent variable. To get the regression line of X on Y, we will have to assume X as the dependent variable and Y as the independent variable. We will then get the two equations for the two regression lines.

The following illustration will illustrate the above points.

Illustration

Plot the regression lines associated with the following data :

Values of X	1	2	3	4	5
Values of Y	166	184	142	180	338

Solution. To obtain the straight line equations of the given values, we will use the following normal equations.

$$\sum(Y) = na + b\,(\sum X)$$

$$\sum(XY) = a\sum X + b(\sum X^2)$$

These equations will give us the values of a and b which we will fit in the equation of the straight line,

$$Y = a + bX$$

This will give us the regression line of Y on X. The value of a in this problem would be a = 100 and b = 34. So the equation would be :

$$Y = 100 + 34X \text{ ..... regression equation of Y and X.}$$

To find the regression equation of X on Y, we have to find the values of a and b in the equation :

$$X = a + bY$$

and in this case, the two normal equations would be :

$$\sum(X) = na + b\,(\sum Y)$$

$$\sum(XY) = a\sum Y + b(\sum Y^2)$$

with these equations the equation, of the straight line for X on Y would be:

$$X = 0.172 + 0.014Y \text{ ..... regression equation of X and Y.}$$

as the value of a from the normal equations would be 0.172 and b = 0.014.

Thus, the equations of straight line or the lines of the best fit would be as follows:

$$Y = 100 + 34X \qquad \text{..... (i)}$$

$$X = 0.172 + 0.014Y \qquad \text{..... (ii)}$$

From the first equation, we will find any two values of Y for some values of X and plot them on the graph paper to get the Regression Line of Y on X.

From the second equation. we will find any two values of X for some values of Y and plot them on the graph paper to get the Regression Line on X on Y.

The two lines would cut each other at the point which gives, the average values of X and Y.

The following graph would emerge from these lines :

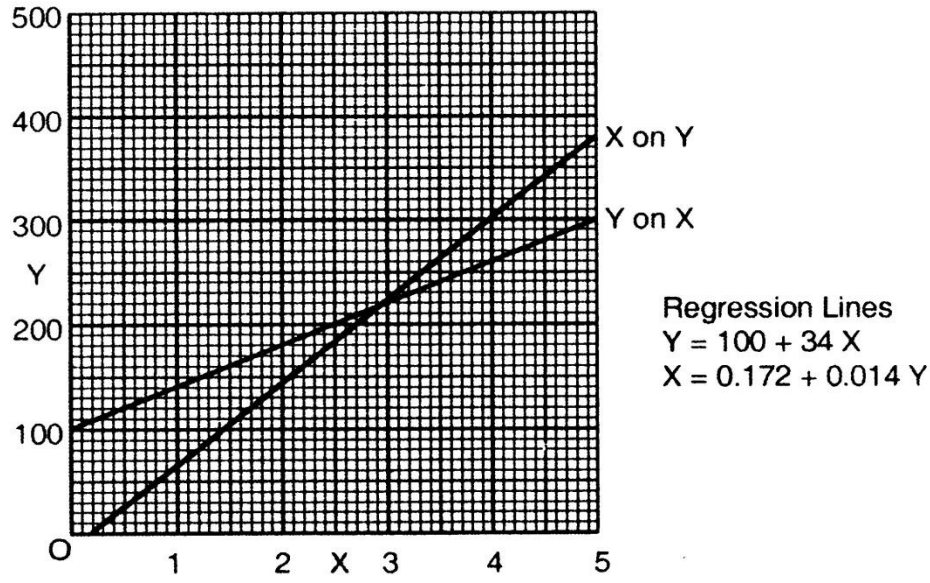


Fig. 1

It would be observed from the above graph that the two regression lines cut each other at the point of Arithmetic mean of the two series. If we have to find out any value of Y for a given value of X, we will draw a perpendicular from the X series (for the given value of X) and the point at which it cuts the regression line of Y on X will indicate the coupled value of Y which can be read on the Y-scale (by drawing a line parallel to X axis from the point where the perpendicular joins the regression line of Yon Y). Thus, when X = 2, Y would be 168. Similarly, values of Y series can be found by using the regression line of X on Y.

As was pointed out earlier, the regression line or the line of the best fit is one from which the square of the deviations between the given values of the variable and its estimated values is the least. In our problem, the Y series is on the Vertical scale and so the square of the deviations (measured on vertical scale) between' the original figures and the regression line would be the least. For the X series, the sum of squares on the horizontal scale would be the least. The following two graphs illustrate these points :

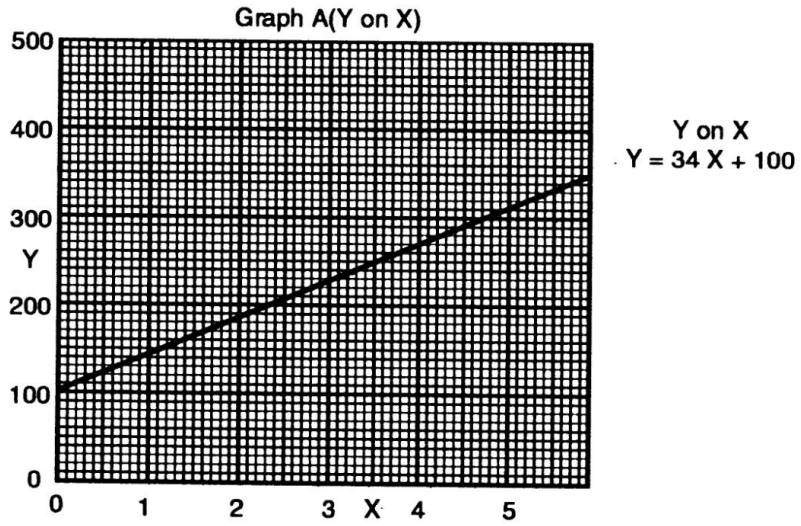


Fig. 2

In the above graph, besides the regression line of Y on X, we have plotted the original figures of Y and the perpendiculars show the deviation between original and computed figures. The sum of the square of these deviations would be the least when deviations are taken from the regression line values.

Fig. 3 shows similar deviations of the original values of X series with the values in the regression equation of X on Y.

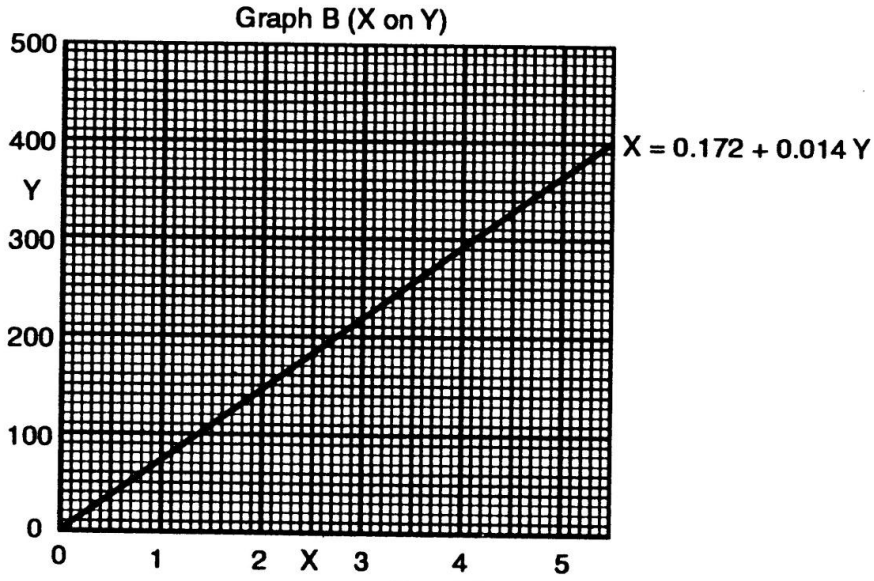


Fig. 3

From the graph, it is clear that the deviations of X series from the regression line are the least and so the sum of their squares would also be the least. In fact, the sum of the deviations is always zero and so the squares of the deviations have the least value.

Why Two Regression Lines

Very often, this question is asked as to why there should be two regression lines to obtain the values of Y and x : and why one regression line does not serve the purpose. The answer is simple and it is that one regression line cannot minimise the sum of squares of deviations for both the X and Y series unless the relationship between them indicates perfect positive or negative correlation. In case of perfect correlations, one regression line is enough because X and Y series have the same type of deviations. Ordinarily, in social sciences, perfect correlation is very rarely found. For this reason, one regression line minimises the sum of the squares of deviations of the X series and the other regression line takes care of the deviations of Y series. This point is obvious from Figs. 2 and 3. Fig. 2 has minimized the sum of the squares of deviations of Y series and Fig. 3, of the X series. In the first case, the deviations have been measured on the vertical scale and in the other on horizontal scale.

19.10. REGRESSION EQUATIONS

Regression equations are algebraic expressions of the regression lines. Since there are two regression lines, there are two regression equations. Regression line of X on Y give the best possible mean values of X for given values of Y and, similarly, the regression line of Y on X gives the best possible mean values of Y for given values or X As such, regression equation of X on Y would be used to

describe the variation in the values of X for given changes in the values of Y and, similarly, the regression equation of Y on X would be used to describe the variation in the value of Y for given changes in the values of X.

The regression equation of X on Y is X = a + bY and the regression equation of Y on X is Y = a + bX. These are the equations of a straight line. In these equations, the values of a and b are constant which determine the positions of the lines of regression. The parameter a indicates the level of the line of regression (the distance of the line above or below the origin). The parameter b determines the slope of the line, i.e., the corresponding change in X in relation to per unit change in Y or vice versa.

The values of a and b in the above equations are found by the Method of Least Squares-reference to which was made earlier. The values of a and b are found with the help of normal equations given below :

(i) $\sum Y = na + b \sum X$

(ii) $\sum X = na + b \sum Y$

$\sum XY = a \sum X + b \sum X^2$ $\sum XY = a \sum Y + b \sum Y^2$

We shall take an example to illustrate this technique of finding out the values of a and b and with their help to obtain the regression equations.

Illustration

From the following data, obtain the two regression equations using the method of Least Squares.

X	2	4	6	8	10
Y	5	7	9	8	11

Solution

Computation of Regression Equations

X	**Y**	**XY**	**X^2**	**Y^2**
2	5	10	4	25
4	7	28	16	49
6	9	54	36	81
8	8	64	64	64
10	11	110	100	121
$\sum X = 30$	**$\sum Y = 40$**	**$\sum XY = 266$**	**$\sum X^2 = 220$**	**$\sum Y^2 = 340$**

Regression Equation Y on X is of the form Y = a + bX.

To find the values of a and b, the following two normal equations are used:

$$\sum Y = na + b \sum X$$

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

Substituting the values from the table we get:

$40 = 5a + 30b$ ..... (i)

$266 = 30a + 220b$ ..... (ii)

Multiplying equation (i) by 6 we get:

$240 – 30a + 180b$ ..... (iii)

Subtracting equation (iii) from equation (ii) we get:

$40\ b = 26$ or $b = 0.65$

Substituting the value of b in equation (i) we get:

$40 = 5a + 19.5$ or $5a = 20.5 \Rightarrow a = 4.1$

Substituting the values of a and b in the regression equation Yon X we get:

$Y = 4.1 + 0.65\ X \rightarrow$ This is the Regression Equation of Y on X

Regression Equation of X on Y is of the form $X = a + bY$

The two normal equation are :

$$\sum X = na + b\sum Y$$

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

Substituting the values in these equations we get:

$30 = 5a + 40b$ ..... (i)

$266 = 40a + 340b$ ..... (ii)

Multiplying equation (i) by 8 we get:

$240 = 40a + 320b$ ..... (iii)

Subtracting (iii) from (ii) we :

$20b = 26 \Rightarrow b = 1.3$

Substituting the value of b in equation (i) we get:

$30 = 5a + 52 \Rightarrow 5a = – 22$

or $a = – 4.4$

Substituting the values of a and b in the recession equation X on Y we get:

$X = – 4.4 + 1.3\ Y$. This is the regression equation of X on Y.

19.11. Comparison of Correlation and Regression Methods

Both the correlation and regression analysis helps us in studying the relation-ship between two

variables yet they differ in their approach, and objectives.

1. Correlation studies are meant for studying the Co-variation of the two variables. They tell us whether the variables under study move in the same direction or in reverse directions. The degree of their co-variation is also reflected in the correlation coefficient, but the correlation study does not study the *nature of relationship.* It does not tell us about the relative movement in the variables under study and we cannot predict the value of one variable by taking into account the value of the other variable. This is possible through regression analysis, i.e., regression analysis can be used for prediction where correlation cannot be used for prediction.

2. Correlation between two series is not necessarily a cause and effect relationship. A high degree of positive correlation between price and supply does not mean that, supply is the effect of prices. There may be no cause and effect relationship between the variables under study and yet they may be correlated. Regression on the other hand presumes one variable as a cause and the other as its effect. The independent variable is supposed to be affecting the dependent variable and as such we can estimate the values of the dependent variable for a given value of the independent variable.

3. The Coefficient of correlation r varies between ± 1 i.e., $-1 \leq r \leq 1$. The regression coefficients have the same sings as the correlation coefficient. If r is positive regression coefficient would also be positive and if r is negative the regression coefficients would also be negative.

4. Further, whereas correlation coefficient cannot exceed unity, but one of the regression coefficients can have a value higher than unity but the product of the two regression coefficients can never exceed unity because correlation coefficient is the square root of the product of the two regression coefficients.

19.12. SUMMARY

The term correlation indicates the relationship between two such variables in which with change of values of one variable the values of the other variable change if the two variable work together. Thus, correlation analysis attempts to determine the degree of relationship between variables.

Regression is the measure of the average relationship between two or more variables in terms of the original units of the data.

19.13. FURTHER STUDY

1. Sankalp Gaurav, Business Statistics, Agra book International , Agra
2. Sinha, V.C., Principles of Statistics.
3. Gupta, S.B., Principles of Statistics.
4. Monga, G.S., Elementary Statistics.

Unit 20 INDEX NUMBERS

UNIT STRUCTURE

OBJECTIVES

After going through this unit you have a knowledge of

- Index number and its application
- Types of index number
- Characteristics features and relevance of index number
- How to calculate index number
- Test to satisfy good index number

20.1 MEANING OF INDEX NUMBERS

Index numbers are intended to measure the degree of economic changes over time. These numbers are values stated as a percentage of a single base figure. Index numbers are important in economic statistics. In simple terms, an index (or index number) is a number displaying the level of a variable relative to its level (set equal to 100) in a given base period.

Index numbers are intended to study the change in the effects of such factors which cannot be measured directly. Bowley stated that "Index numbers are used to gauge the changes in some quantity which we cannot observe directly". It can be explained through example in which changes in business activity in a nation are not capable of direct measurement but it is possible to study relative changes in business activity by studying the variations in the values of some such factors which affect business activity, and which are proficient of direct measurement.

Index numbers are usually applied in statistical device to measure the combined fluctuations in a group related variables. If statistician or researcher wants to compare the price level of consumer items today with that predominant ten years ago, they are not interested in comparing the prices of

only one item, but in comparing some sort of average price levels (Srivastava, 1989). With the support of index numbers, the average price of several articles in one year may be compared with the average price of the same quantity of the same articles in a number of different years. There are several sources of 'official' statistics that contain index numbers for quantities such as food prices, clothing prices, housing, and wages.

Index numbers may be categorized in terms of the variables that they are planned to measure. In business, different groups of variables in the measurement of which index number techniques are normally used are price, quantity, value, and business activity.

20.2 IMPORTANCE OF INDEX NUMBERS:

Index numbers are used to measure all types of quantitative changes in different fields.

Various advantages of index numbers are given below:

1. General Importance: In general, index numbers are very useful in a number of ways:

(a) They measure changes in one variable or in a group of variables.

(b) They are useful in making comparisons with respect to different places or different periods of time,

(c) They are helpful in simplifying the complex facts.

(d) They are helpful in forecasting about the future,

(e) They are very useful in academic as well as practical research.

2. Measurement of Value of Money:

Index numbers are used to measure changes in the value of money or the price level from time to time. Changes in the price level generally influence production and employment of the country as well as various sections of the society. The price index numbers also forewarn about the future inflationary tendencies and in this way, enable the government to take appropriate anti- inflationary measures.

3. Changes in Cost of Living:

Index numbers highlight changes in the cost of living in the country. They indicate whether the cost of living of the people is rising or falling. On the basis of this information, the wages of the workers can be adjusted accordingly to save the wage earners from the hardships of inflation.

4. Changes in Production:

Index numbers are also useful in providing information regarding production trends in different sectors of the economy. They help in assessing the actual condition of different industries, i.e., whether production in a particular industry is increasing or decreasing or is constant.

5. Importance in Trade:

Importance in trade with the help of index numbers, knowledge about the trade conditions and trade trends can be obtained. The import and export indices show whether foreign trade of the country is increasing or decreasing and whether the balance of trade is favourable or unfavourable.

6. Formation of Economic Policy:

Index numbers prove very useful to the government in formulating as well as evaluating economic policies. Index numbers measure changes in the economic conditions and, with this information, help the planners to formulate appropriate economic policies. Further, whether particular economic policy is good or bad is also judged by index numbers.

7. Useful in All Fields:

Index numbers are useful in almost all the fields. They are specially important in economic field.

Some of the specific uses of index numbers in the economic field are:

(a) They are useful in analysing markets for specific commodities.

(b) In the share market, the index numbers can provide data about the trends in the share prices,

(c) With the help of index numbers, the Railways can get information about the changes in goods traffic.

(d) The bankers can get information about the changes in deposits by means of index numbers.

20.3 LIMITATIONS OF INDEX NUMBERS:

Index number technique itself has certain limitations which have greatly reduced its usefulness:

(i) Because of the various practical difficulties involved in their computation, the index numbers are never cent per cent correct.

(ii) There are no all-purpose index numbers. The index numbers prepared for one purpose cannot be used for another purpose. For example, the cost-of-living index numbers of factory workers cannot be used to measure changes in the value of money of the middle income group.

(iii) Index numbers cannot be reliably used to make international comparisons. Different countries include different items with different qualities and use different base years in constructing index numbers.

(iv) Index numbers measure only average change and indicate only broad trends. They do not provide accurate information.

(v) While preparing index numbers, quality of items is not considered. It may be possible that a general rise in the index is due to an improvement in the quality of a product and not because of a rise in its price.

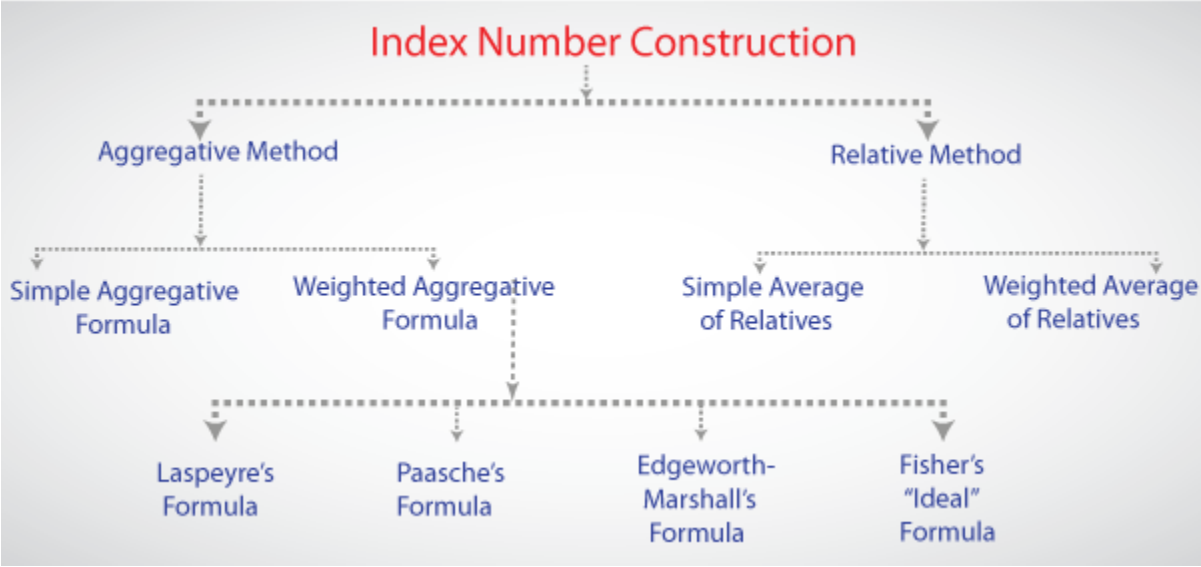
20.4 TYPES OF INDEX NUMBERS

Simple Index Number: A simple index number is a number that measures a relative change in a single variable with respect to a base. These type of Index numbers are constructed from a single item only.
Composite Index Number: A composite index number is a number that measures an average relative changes in a group of relative variables with respect to a base. A composite index number is built from changes in a number of different items.

20.4.1 Price index Numbers: Price index numbers measure the relative changes in prices of a commodity between two periods. Prices can be either retail or wholesale. Price index number are useful to comprehend and interpret varying economic and business conditions over time.

20.4.2 Quantity Index Numbers: These types of index numbers are considered to measure changes in the physical quantity of goods produced, consumed or sold of an item or a group of items. Methods of constructing index numbers: There are two methods to construct index numbers: Price relative and aggregate methods

Srivastava, 1989).



In aggregate methods, the aggregate price of all items in a given year is expressed as a percentage of same in the base year, giving the index number.

$$\text{Index Numbers} = \frac{\text{Aggregate price in the given year}}{\text{Aggregate price in the base year}} \times 100$$

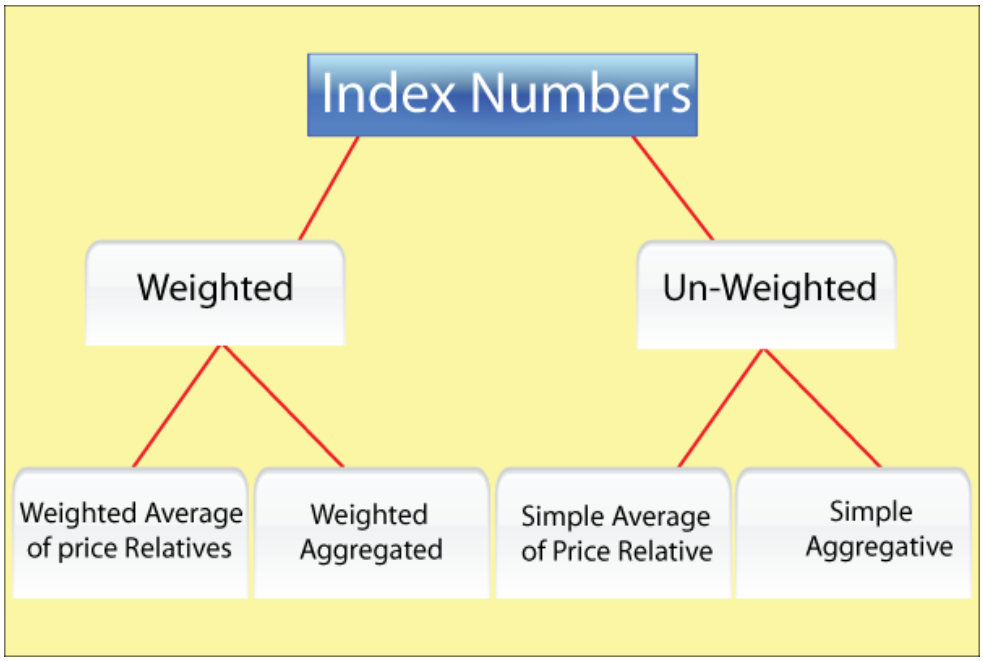
Relative method: The price of each item in the current year is expressed as a percentage of price in base year. This is called price relative and expressed as following formula:

$$\text{Price Relative} = \frac{\text{Price in the given year } (Pn)}{\text{Price in the base year } (P0)} \times 100$$

In simple average of relative method, the current year price is expressed as a price relative of the base year price. These price relatives are then averaged to get the index number. The average used could be arithmetic mean, geometric mean or even median.

Weighted index numbers: These are those index numbers in which rational weights are assigned to various chains in an explicit fashion.

Weighted aggregative index numbers: These index numbers are the simple aggregative type with the fundamental difference that weights are assigned to the various items included in the index.



20.5 Characteristics of index numbers:

1. Index numbers are specialised averages.
2. Index numbers measure the change in the level of a phenomenon.
3. Index numbers measure the effect of changes over a period of time.

Uses of Index number: Index numbers has practical significance in measuring changes in the cost of living, production trends, trade, and income variations. Index numbers are used to measure changes in the value of money. A study of the rise or fall in the value of money is essential for determining the direction of production and employment to facilitate future payments and to know changes in the real income of different groups of people at different places and times. Crowther designated, "By using the technical device of an index number, it is thus possible to measure changes in different aspects of the value of money, each particular aspect being relevant to a different purpose." Basically, index numbers are applied to frame appropriate policies. They reveal trends and tendencies and Index numbers are beneficial in deflating.

The value of money does not remain constant over time. It rises or falls and is inversely related to the changes in the price level. A rise in the price level means a fall in the value of money and a fall in the price level means a rise in the value of money. Thus, changes in the value of money are reflected by the changes in the general level of prices over a period of time. Changes in the general level of prices can be measured by a statistical device known as ‘index number.’

Index number is a technique of measuring changes in a variable or group of variables with respect to time, geographical location or other characteristics. There can be various types of index numbers, but, in the present context, we are concerned with price index numbers, which measures changes in the general price level (or in the value of money) over a period of time.

Price index number indicates the average of changes in the prices of representative commodities at one time in comparison with that at some other time taken as the base period. According to L.V. Lester, “An index number of prices is a figure showing the height of average prices at one time relative to their height at some other time which is taken as the base period.”

20.6 FEATURES OF INDEX NUMBERS:

The following are the main features of index numbers:

(i) Index numbers are a special type of average. Whereas mean, median and mode measure the absolute changes and are used to compare only those series which are expressed in the same units, the technique of index numbers is used to measure the relative changes in the level of a phenomenon where the measurement of absolute change is not possible and the series are expressed in different types of items.

(ii) Index numbers are meant to study the changes in the effects of such factors which cannot be measured directly. For example, the general price level is an imaginary concept and is not capable of direct measurement. But, through the technique of index numbers, it is possible to have an idea of relative changes in the general level of prices by measuring relative changes in the price level of different commodities.

(iii) The technique of index numbers measures changes in one variable or group of related variables. For example, one variable can be the price of wheat, and group of variables can be the price of sugar, the price of milk and the price of rice.

(iv) The technique of index numbers is used to compare the levels of a phenomenon on a certain date with its level on some previous date (e.g., the price level in 2020 as compared to that in 2010 taken as the base year) or the levels of a phenomenon at different places on the same date (e.g., the price level in India in 2020 in comparison with that in other countries in 2020).

20.7 STEPS OR PROBLEMS IN THE CONSTRUCTION OF PRICE INDEX NUMBERS

The construction of the price index numbers involves the following steps or problems:

1. Selection of Base Year:

The first step or the problem in preparing the index numbers is the selection of the base year. The base year is defined as that year with reference to which the price changes in other years are compared and expressed as percentages. The base year should be a normal year.

In other words, it should be free from abnormal conditions like wars, famines, floods, political instability, etc. Base year can be selected in two ways- (a) through fixed base method in which the base year remains fixed; and (b) through chain base method in which the base year goes on changing, e.g., for 2020 the base year will be 2019, for 2019 it will be 2018, and so on.

2. Selection of Commodities:

The second problem in the construction of index numbers is the selection of the commodities. Since all commodities cannot be included, only representative commodities should be selected keeping in view the purpose and type of the index number.

In selecting items, the following points are to be kept in mind:

(a) The items should be representative of the tastes, habits and customs of the people.

(b) Items should be recognizable,

(c) Items should be stable in quality over two different periods and places.

(d) The economic and social importance of various items should be considered

(e) The items should be fairly large in number.

(f) All those varieties of a commodity which are in common use and are stable in character should be included.

3. Collection of Prices:

After selecting the commodities, the next problem is regarding the collection of their prices:

(a) From where the prices to be collected;

(b) Whether to choose wholesale prices or retail prices;

(c) Whether to include taxes in the prices or not etc.

While collecting prices, the following points are to be noted:

(a) Prices are to be collected from those places where a particular commodity is traded in large quantities.

(b) Published information regarding the prices should also be utilised,

(c) In selecting individuals and institutions who would supply price quotations, care should be taken that they are not biased.

(d) Selection of wholesale or retail prices depends upon the type of index number to be prepared. Wholesale prices are used in the construction of general price index and retail prices are used in the construction of cost-of-living index number.

(e) Prices collected from various places should be averaged.

4. Selection of Average:

Since the index numbers are, a specialised average, the fourth problem is to choose a suitable average. Theoretically, geometric mean is the best for this purpose. But, in practice, arithmetic mean is used because it is easier to follow.

5. Selection of Weights:

Generally, all the commodities included in the construction' of index numbers are not of equal importance. Therefore, if the index numbers are to be representative, proper weights should be assigned to the commodities according to their relative importance.

For example, the prices of books will be given more weightage while preparing the cost-of-living index for teachers than while preparing the cost-of-living index for the workers. Weights should be unbiased and be rationally and not arbitrarily selected.

6. Purpose of Index Numbers:

The most important consideration in the construction of the index numbers is the objective of the index numbers. All other problems or steps are to be viewed in the light of the purpose for which a particular index number is to be prepared. Since, different index numbers are prepared with specific purposes and no single index number is 'all purpose' index number, it is important to be clear about the purpose of the index number before its construction.

7. Selection of Method:

The selection of a suitable method for the construction of index numbers is the final step.

There are two methods of computing the index numbers:

(a) Simple index number and

(b) Weighted index number.

Simple index number again can be constructed either by – (i) Simple aggregate method, or by (ii) simple average of price relative's method. Similarly, weighted index number can be constructed either by (i) weighted aggregative method, or by (ii) weighted average of price relative's method. The choice of method depends upon the availability of data, degree of accuracy required and the purpose of the study.

20.8 Construction of Price Index Numbers (Formula and Examples):

Construction of price index numbers through various methods can be understood with the help of the following examples:

20.8. 1. Simple Aggregative Method:

In this method, the index number is equal to the sum of prices for the year for which index number is to be found divided by the sum of actual prices for the base year.

The formula for finding the index number through this method is as follows:

$$P_{01} = \frac{\Sigma P_1}{\Sigma P_0} \times 100$$

Where P_{01} Stands for the index number

ΣP_1 Stands for the sum of the prices for the year for which index number is to be found :

ΣP_0 Stands for the sum of prices for the base year.

Commodity	Prices in Base Year 1980 (in Rs.) P_o	Prices in current Year 1988 (in Rs.) P_1
A	10	20
B	15	25
C	40	60
D	25	40
Total	$\Sigma P_0 = 90$	$\Sigma P_1 = 145$

$$\text{Index Number } (P_{01}) = \frac{\Sigma P_1}{\Sigma P_0} \times 100 \; ; \; P_{01} = \frac{145}{90} \times 100 \; ; \; P_{01} = 161.11$$

20.8. 2. Simple Average of Price Relatives Method:

In this method, the index number is equal to the sum of price relatives divided by the number of items and is calculated by using the following formula:

$$P_{01} = \frac{\Sigma R}{N}$$

Where ΣR stands for the sum of price relatives i. e. $R = \frac{P_1}{P_0} \times 100$ and

N stands for the number of items.

Example

Commodity P_0	Base Year Prices (in Rs.) P_1	Current year Prices (in Rs.)	Price Relatives $R = \frac{P_1}{P_0} \times 100$
A	10	20	$\frac{20}{10} \times 100 = 200.0$
B	15	25	$\frac{25}{15} \times 100 = 166.7$
C	40	60	$\frac{60}{40} \times 100 = 150.00$
D	25	40	$\frac{40}{25} \times 100 = 160.0$
N = 4			$\Sigma R = 676.7$

$$\text{Index Number } (p_{01}) = \frac{\Sigma R}{N}$$

$$P_{01} = \frac{676.7}{4}; P_{01} = 169.2$$

20.8.3. Weighted Aggregative Method:

In this method, different weights are assigned to the items according to their relative importance. Weights used are the quantity weights. Many formulae have been developed to estimate index numbers on the basis of quantity weights.

Some of them are explained below:

(*i*) **Laspeyre's Formula.** In this formula, the quantities of base year are accepted as weights.

$$P_{01} = \frac{\Sigma P_1 q_0}{\Sigma P_0 q_0} \times 100$$

Where P_1 is the price in the current year ; P_0 is the price in the base year ; and q_0 is the quantity in the base year.

(*ii*) **Paasche's Formula.** In this formula, the quantities of the current year are accepted as weights.

$$P_{01} = \frac{\Sigma P_1 q_1}{\Sigma P_0 q_1} \times 100$$

Where q_1 is the quantity in the current year.

(*iii*) **Dorbish and Bowley's Formula.** Dorbish and Bowley's formula for estimating weighted index number is as follows :

$$P_{01} = \frac{\frac{\Sigma P_1 q_0}{\Sigma P_0 q_0} + \frac{\Sigma P_1 q_1}{\Sigma P_0 q_1}}{2} \times 100 \quad \text{or} \quad p_{01} = \frac{L + P}{2}$$

Where L is Laspeyre's index and P is paasche's Index.

(*iv*) **Fisher's Ideal Formula.** In this formula, the geometric mean of two indices (i.e., Laspeyre's Index and paasche's Index) is taken :

$$p_{01} = \sqrt{\frac{\Sigma P_1 q_0}{\Sigma P_0 q_0} \times \frac{\Sigma P_1 q_1}{\Sigma P_0 q_1}} \times 100 \quad \text{or} \quad P_{01} = \sqrt{L \times P} \times 100$$

where L is Lespeyre's Index and P is paasche's Index.

Example

Comm-odity	Base Year		Current Year		P_0q_0	P_1q_0	P_0q_1	P_1q_1
	P_0	q_0	P_1	q_1				
A	10	5	20	2	50	100	20	40
B	15	4	25	8	60	100	120	200
C	40	2	60	6	80	120	240	360
D	25	3	40	4	75	120	100	160
Total					265 ΣP_0q_0	440 ΣP_1q_0	480 ΣP_0q_1	760 ΣP_1q_1

(*i*) Laspeyre's Formula :

$$p_{01} = \frac{\Sigma P_1 q_0}{\Sigma P_0 q_0} \times 100$$

$$p_{01} = \frac{440}{265} \times 100 = 166.04$$

(*ii*) Paasche' Formula :

$$p_{01} = \frac{\Sigma P_1 q_1}{\Sigma P_0 q_1} \times 100$$

$$p_{01} = \frac{700}{480} \times 100 = 158.3$$

(*iii*) Dorbish and Bowley's Formula :

$$p_{01} = \frac{\frac{\Sigma P_1 q_0}{\Sigma P_0 q_0} + \frac{\Sigma P_1 q_1}{\Sigma P_0 q_1}}{2} \times 100 = 162.2$$

$$p_{01} = \frac{\frac{440}{265} + \frac{760}{480}}{2} \times 100 = 162$$

(*iv*) Fisher's Ideal Formula :

$$p_{01} = \sqrt{\frac{\Sigma P_1 q_0}{\Sigma P_0 q_0} \times \frac{\Sigma P_1 q_1}{\Sigma P_0 q_1}} \times 100$$

$$p_{01} = \sqrt{\frac{440}{265} \times \frac{760}{480}} \times 100 = 162.1$$

20.8.4. Weighted Average of Relatives Method:

In this method also different weights are used for the items according to their relative importance.

The price index number is found out with the help of the following formula:

$$P_{01} = \frac{\Sigma RW}{\Sigma W}$$

where ΣW stands for the sum of weights of different commodities :
and ΣR stands for the sum of price relatives.

Commodity	Weights W	Base Prices Year P_0	Current Year Prices P_1	Price Relatives $R = \frac{P_1}{P_0} \times 100$	RW
A	5	10	20	20/10 × 100 = 200.0	1000.0
B	4	15	25	25/15 × 100 = 166.7	666.8
C	2	40	60	60/40 ×100 = 150.0	300.0
D	3	25	40	40/25 × 100 = 160.0	480.0
Total	ΣW=14				ΣRW = 2446.8

$$\text{Index Number } (P_{01}) = \frac{\Sigma RW}{\Sigma W}$$

$$p_{01} = \frac{2446.8}{14} = 174.8$$

20.9 TEST TO BE SATISFIED BY A GOOD INDEX NUMBER

We have observed that there are many Index Numbers. Now the question is which one is better. To judge this we check whether the index numbers satisfies the following mathematical tests.

1. The Commodity reversal test

2. Unit test

3. Time reversal test

4. Circular test

5. Factor reversal test

1. **The Commodity reversal test** : An index number is said to satisfy this test if it remains unchanged even if the order in which the commodities are considered is changed. All the index numbers satisfies this test.

2. **Unit test :** If the index number is independent of the units in which the prices and quantities are expressed it is said to satisfy the unit test. All index numbers considered so far satisfies this test.

3. **Time reversal test** : Let I0k denote the index number calculated with the period denoted by '0' as the base period and the period denoted by 'k' as the current and Ik0 the index number calculated with the periods interchanged. The index number is said to satisfy the time reversal test if I0k x Ik0 = 1. Only the simple G.M. index number and Fishers Ideal Index number satisfies this test.

4. **Circular test:** This is another test for the adequacy of an index number. This test is based on the suitability of the base and is an extension of the time reversal test. Let 0, 1 and 2 are three years and I01, I12 and I02 are the indices for year 1 with 0 as base year, year 2 with 1 as base year and year 2 with 0 as base respectively. The circular test is said to be satisfied if, I01 x I12 = I02 or I01 x I12 x I02 =1. The index number considered so far only simple G.M. index number, aggregate index number and Fishers ideal number satisfies this test.

5. Factor reversal test This test is applicable only to weighted index number. Let Ipq be the index number calculated with p denoting the price and q denoting the quantity and Iqp denote the index number obtained by interchanging p and q. The index number is said to satisfy this test if Ipq x Iqp = $\sum \sum$ p0 q0 / pk qk . Only Fisher's ideal index number satisfies this test.

20.10 TYPES OF INDEX NUMBERS:

There are different types of index numbers some i**mportant types of index numbers are discussed below:**

1. Wholesale Price Index Numbers:

Wholesale price index numbers are constructed on the basis of the wholesale prices of certain important commodities. The commodities included in preparing these index numbers are mainly raw-materials and semi-finished goods. Only the most important and most price-sensitive and semi-finished goods which are bought and sold in the wholesale market are selected and weights are assigned in accordance with their relative importance.

The wholesale price index numbers are generally used to measure changes in the value of money. The main problem with these index numbers is that they include only the wholesale prices of raw materials and semi-finished goods and do not take into consideration the retail prices of goods and services generally consumed by the common man. Hence, the wholesale price index numbers do not reflect true and accurate changes in the value of money.

2. Retail Price Index Numbers:

These index numbers are prepared to measure the change in the value of money on the basis of the retail prices of final consumption goods. The main difficulty with this index number is that the retail

price for the same goods and for continuous periods is not available. The retail prices represent larger and more frequent fluctuations as compared to the wholesale prices.

3. Cost-of-Living Index Numbers:

These index numbers are constructed with reference to the important goods and services which are consumed by common people. Since the number of these goods and services is very large, only representative items which form the consumption pattern of the people are included. These index numbers are used to measure changes in the cost of living of the general public.

4. Working Class Cost-of-Living Index Numbers:

The working class cost-of-living index numbers aim at measuring changes in the cost of living of workers. These index numbers are consumed on the basis of only those goods and services which are generally consumed by the working class. The prices of these goods and index numbers are of great importance to the workers because their wages are adjusted according to these indices.

5. Wage Index Numbers:

The purpose of these index numbers is to measure time to time changes in money wages. These index numbers, when compared with the working class cost-of-living index numbers, provide information regarding the changes in the real wages of the workers.

6. Industrial Index Numbers:

Industrial index numbers are constructed with an objective of measuring changes in the industrial production. The production data of various industries are included in preparing these index numbers.

20.11 SOLVED EXAMPLES

Question : Taking 2004 as base year, construct the index numbers of the years 2005 and 2009.

Year	2004	2005	2006	2007	2008	2009
Price	10	14	16	20	22	24

Solution

Year	Price
2004	10
2005	14
2006	16
2007	20
2008	22
2009	24

Given:

$P_0 = 10$

Index number for year 2005:

$$P_{01} = \frac{P_1}{P_0} \times 100 = \frac{14}{10} \times 100$$

$$\therefore \boxed{P_{01} = 140}$$

Index number for year 2009:

$$P_{01} = \frac{24}{10} \times 100$$

$$\therefore \boxed{P_{01} = 240}$$

Question : Construct index number by Price Relative Method taking 2004 as base year.

Price per Unit in Rs

Year	A	B	C	D
2004	25	18	16	21
2012	20	22	24	22
2013	25	20	25	25
2014	28	24	30	26

Solution

	2004 (P_0)	2012 (P_1)	Price Relative $= \frac{P_1}{P_0} \times 100$
A	25	20	$\frac{20}{25} \times 100 = 80$
B	18	22	$\frac{22}{18} \times 100 = 122.22$
C	16	24	$\frac{24}{16} \times 100 = 150$
D	21	22	$\frac{22}{21} \times 100 = 104.76$
Total			456.98

$$P_{01} = \frac{\sum\left(\frac{P_1}{P_0} \times 100\right)}{N}$$

$$P_{01} = \frac{456.98}{4}$$

$$\therefore \boxed{P_{01} = 114.245}$$

	2004 (P_0)	2013 (P_1)	Price Relative $= \frac{P_1}{P_0} \times 100$
A	25	25	$\frac{25}{25} \times 100 = 100$
B	18	20	$\frac{20}{18} \times 100 = 111.11$
C	16	25	$\frac{25}{16} \times 100 = 156.25$
D	21	25	$\frac{25}{21} \times 100 = 119.04$
Total			486.4

$$P_{01} = \frac{486.4}{4}$$

$$\therefore \boxed{P_{01} = 121.60}$$

	2004 (P_0)	2014 (P_1)	Price Relative $= \frac{P_1}{P_0} \times 100$
A	25	28	$\frac{28}{25} \times 100 = 280$
B	18	24	$\frac{24}{18} \times 100 = 133.33$
C	16	30	$\frac{30}{16} \times 100 = 187.5$
D	21	26	$\frac{26}{21} \times 100 = 123.80$
Total			556.63

$$P_{01} = \frac{556.63}{4}$$

$$\therefore \boxed{P_{01} = 139.16}$$

Question : Find out the index number of the following data with Laspeyre's Method :

Commodity	2013		2014	
	Price	**Quantity**	**Price**	**Quantity**
A	70	7	80	6
B	62	3	74	2

Solution

	P_0	q_0	P_0q_0	P_1	q_1	P_1q_0
A	70	7	490	80	6	560
B	62	3	186	74	2	222
			$\Sigma P_0q_0 = 676$			$\Sigma P_1q_0 = 782$

Laspeyre's Price index

$$P_{01} = \frac{\sum P_1q_0}{\sum P_0q_0} \times 100$$

$$P_{01} = \frac{782}{676} \times 100$$

$$\therefore \boxed{P_{01} = 115.68}$$

Question : Construct index numbers of the following data with Laspeyre's and Paasche's Methods :

Commodity	**Base Year**		**Current Year**	
	Price	**Quantity**	**Price**	**Quantity**
A	10	0.80	11	0.70
B	8	0.85	9	0.90
C	5	1.30	5.5	0.80

Solution

	p_0	q_0	p_0q_0	p_1	q_1	p_1q_1	p_0q_1	p_1q_0
A	0.8	10	8	0.7	11	7.7	8.8	7
B	0.85	8	6.8	0.9	9	8.1	7.65	7.2
C	1.35	5	6.5	0.8	5.5	4.4	7.15	4
			$\Sigma p_0q_0 = 21.3$			$\Sigma p_1q_1 = 20.2$	$\Sigma p_0q_1 = 23.60$	$\Sigma p_0q_1 = 18.2$

Laspeyre's price index:

$$P_{01} = \frac{\sum P_1 q_0}{\sum P_0 q_0} \times 100$$

$$P_{01} = \frac{18.2}{21.3} \times 100$$

$$\therefore \boxed{P_{01} = 85.45}$$

Paasche's price index:

$$P_{01} = \frac{\sum P_1 q_1}{\sum P_0 q_1} \times 100$$

$$P_{01} = \frac{20.2}{23.60} \times 100$$

$$\therefore \boxed{P_{01} = 85.59}$$

Question : Construct index numbers of the following data by Fisher's Method :

Commodity	**Base Year**		**Current Year**	
	Price	**Value**	**Price**	**Value**
A	3	18	7	14
B	5	35	10	100
C	6	42	11	55
D	4	32	6	60
E	8	24	9	36

Solution

	P_0	Base year value	$q_0=\frac{\text{value}}{P_0}$	P_0q_0	P_1	Current year value	$q_1=\frac{\text{value}}{P_1}$	P_1q_1	P_1q_0	P_0q_1
A	3	18	6	18	7	14	2	14	42	6
B	5	35	7	35	10	100	10	100	70	50
C	6	42	7	42	11	55	5	55	77	30
D	4	32	8	32	6	60	10	60	48	40
E	8	24	3	24	9	36	4	36	27	32
				$\sum P_0q_0=151$				$\sum P_1q_1=265$	$\sum P_1q_0=264$	$\sum P_0q_1=158$

Fisher'sPrice Index:

$$P_{01}=\sqrt{\frac{\sum P_1q_0}{\sum P_0q_0}\times\frac{\sum P_1q_1}{\sum P_0q_1}}\times 100$$

$$P_{01}=\sqrt{\frac{264}{151}\times\frac{265}{158}}\times 100$$

$$\therefore \boxed{P_{01}=171.24}$$

Question: Calculate the Laspeyre's, Paasche's and Fisher's price index number for the following data. Interpret on the data.

Commodities	Price		Quantity	
	2000	2010	2000	2010
Rice	38	35	6	7
Wheat	12	18	7	10
Rent	10	15	10	15
Fuel	25	30	12	16
Miscellaneous	30	33	8	10

Solution

Commodities	Price		Quantity		p_0q_0	p_0q_1	p_1q_0	p_1q_1
	2000 (p_0)	2010 (p_1)	2000 (q_0)	2010 (q_1)				
Rice	38	35	6	7	228	266	210	245
Wheat	12	18	7	10	84	120	126	180
Rent	10	15	10	15	100	150	150	225
Fuel	25	30	12	16	300	400	360	480
Miscellaneous	30	33	8	10	240	300	264	330
				Total	**952**	**1236**	**1110**	**1460**

Laspeyre's price index number

$$P_{01}^{L}=\frac{\sum p_1q_0}{\sum p_0q_0}\times 100=\frac{1110}{952}\times 100=116.60$$

On an average, there is an increase of 16.60 % in the price of the commodities when the year 2000 compared with the year 2010.

Paasche's price index number

$$P_{01}^{P} = \frac{\sum p_1 q_1}{\sum p_0 q_1} \times 100 = \frac{1460}{1236} \times 100 = 118.12$$

On an average, there is an increase of 18.12 % in the price of the commodities when the year 2000 compared with the year 2010.

Fisher's price index number

$$P_{01}^{F} = \sqrt{\frac{\sum p_1 q_0 \times \sum p_1 q_1}{\sum p_0 q_0 \times \sum p_0 q_1}} \times 100 = \sqrt{\frac{1110 \times 1460}{952 \times 1236}} \times 100 = 117.36$$

On an average, there is an increase of 17.36 % in the price of the commodities when the year 2000 compared with the year 2010.

20.12 Further Readings

5. Sankalp Gaurav, Business Statistics, Agra Book International
6. Alhance D.N. and Alhance Beena, Fundamental of Statistics.
7. Monga, G.S., Elementary Statistics.
8. Gupta, S.B., Principles of Statistics.

UNIT 21 : STATISTICAL QUALITY CONTROL (SQC)

UNIT STRUCTURE

21.0 OBJECTIVES

After Going through this unit you are able to understand

- About statistical quality control
- SQC and its key components
- Uses of SQC
- Benefits and limitations of SQC

21.1 INTRODUCTION

Statistical Quality Control (SQC) is a methodology used to monitor and control processes to ensure they operate within predetermined limits. It employs statistical methods to detect and prevent defects, reducing waste and improving overall quality.

21.2 KEY COMPONENTS OF SQC

Key components of SQC include:

21.2.1 **Control Charts**: Graphical representations of process data over time, used to detect deviations from normal behaviour.

21.2.2 **Process Capability Analysis**: Evaluates a process's ability to produce output within specified limits.

21.2.3 **Hypothesis Testing**: Statistical tests to determine if a process has changed or if differences exist between processes.

21.2.4 **Design of Experiments (DOE):** A structured approach to identifying the relationship between process variables and output.

21.2.5 **Process Optimization:** Using statistical methods to identify optimal process settings.

Let’s explore the above mention points

21.2.1 Control Charts

Control Charts are a fundamental tool in Statistical Quality Control (SQC). They're graphical representations of process data over time, used to:

1. Monitor process behavior

2. Detect deviations from normal behavior (special causes)

3. Control processes to prevent defects

Key components of Control Charts:

1. Centre Line (CL): Average process value

2. Upper Control Limit (UCL): Maximum acceptable value

3. Lower Control Limit (LCL): Minimum acceptable value

4. Data Points: Individual process measurements plotted over time

Types of Control Charts:

1. $\bar{X}$-R Chart: Monitors process mean ($\bar{X}$) and range (R)

2. $\bar{X}$-s Chart: Monitors process mean ($\bar{X}$) and standard deviation (s)

3. p-Chart: Monitors proportion of defective units

4. c-Chart: Monitors count of defects per unit

5. Individuals-Moving Range (I-MR) Chart: Monitors individual data points and moving range

Interpretation of Control Charts:

1. In-control: Data points within UCL and LCL, indicating a stable process

2. Out-of-control: Data points outside UCL or LCL, indicating a special cause

3. Trends: Gradual changes in process behavior

4. Shifts: Sudden changes in process behavior

By using Control Charts, you can:

1. Improve process stability

2. Reduce variability

3. Detect issues early

4. Optimize process performance

21.2.2 Process Capability Analysis (PCA)

Process Capability Analysis (PCA) is a statistical method used to evaluate a process's ability to produce output within specified limits or tolerances. It helps to:

1. Assess process performance

2. Identify opportunities for improvement

3. Determine if a process is capable of meeting requirements

Key components of PCA:

1. Cp (Capability Index): Measures process spread relative to specification limits

2. Cpk (Cantered Capability Index): Measures process centring and spread

3. Pp (Process Performance Index): Measures process performance over time

4. Ppk (Process Performance Capability Index): Measures process performance and centring

Interpretation of PCA metrics:

1. Cp/Cpk > 1: Process is capable of meeting specifications

2. Cp/Cpk < 1: Process is not capable of meeting specifications

3. Cp/Cpk = 1: Process is barely capable of meeting specifications

Steps for conducting PCA:

1. Collect process data

2. Determine specification limits

3. Calculate Cp, Cpk, Pp, and Ppk

4. Interpret results

5. Improve process as needed

Benefits of PCA:

1. Improved quality

2. Reduced variability

3. Increased customer satisfaction

4. Data-driven decision making

Common applications of PCA:

1. Manufacturing: Evaluating production processes

2. Service: Assessing service quality

3. Healthcare: Improving patient outcomes

21.2.3 HYPOTHESIS TESTING

Hypothesis Testing is a statistical method used to make inferences about a population based on a sample of data. It involves:

1. Formulating a null hypothesis (H0) and an alternative hypothesis (H1)

2. Selecting a significance level (α)

3. Collecting sample data

4. Calculating a test statistic

5. Determining the p-value

6. Making a decision to reject or fail to reject H0

Types of Hypothesis Tests:

1. Two-tailed test: Tests for differences in both directions

2. One-tailed test: Tests for differences in one direction

3. Left-tailed test: Tests for differences in the left tail

4. Right-tailed test: Tests for differences in the right tail

Common Hypothesis Tests:

1. t-test: Compares means of two groups

2. ANOVA: Compares means of three or more groups

3. Chi-squared test: Tests for independence or goodness of fit

4. Regression analysis: Tests for relationships between variables

Interpretation of Hypothesis Test Results:

1. p-value < α: Reject H0 (statistically significant)

2. p-value > α: Fail to reject H0 (not statistically significant)

3. Type I error: Rejecting H0 when it's true (α)

4. Type II error: Failing to reject H0 when it's false (β)

Hypothesis Testing Applications:

1. Comparing treatment effects

2. Evaluating process changes

3. Assessing relationships between variables

4. Validating models or predictions

21.2.4 Design of Experiments (DOE)

Design of Experiments (DOE) is a systematic approach to planning, conducting, and analyzing experiments to:

1. Identify the relationship between variables

2. Optimize processes or products

3. Minimize variability

4. Maximize efficiency

Key components of DOE:

1. Factors: Variables that affect the response

2. Levels: Settings or values of the factors

3. Response: Outcome or variable being measured

4. Experimental Design: Plan for assigning factors to levels

Types of DOE:

1. Full Factorial: All possible combinations of factors and levels

2. Fractional Factorial: Subset of combinations

3. Response Surface Methodology (RSM): Models relationships between factors and response

4. Taguchi Methods: Emphasizes robustness and signal-to-noise ratio

DOE Steps:

1. Define problem and objectives

2. Identify factors and levels

3. Choose experimental design

4. Conduct experiment

5. Analyze data

6. Draw conclusions and make recommendations

Benefits of DOE:

1. Improved understanding of process relationships

2. Optimized processes and products

3. Reduced variability and defects

4. Increased efficiency and productivity

5. Data-driven decision making

Common applications of DOE:

1. Manufacturing: Process optimization and improvement

2. Product Development: Design and testing

3. Quality Control: Defect reduction and reliability

4. Research and Development: Understanding complex relationships

21.2.5 Process Optimization

Process Optimization is the application of methods and tools to improve the efficiency, effectiveness, and quality of processes. It involves:

1. Identifying opportunities for improvement

2. Analyzing process data and behavior

3. Modeling and simulating processes

4. Implementing changes and monitoring results

Goals of Process Optimization:

1. Increase efficiency and productivity
2. Reduce costs and waste
3. Improve quality and customer satisfaction
4. Enhance flexibility and adaptability

Methods and Tools for Process Optimization:

1. Statistical Process Control (SPC)
2. Design of Experiments (DOE)
3. Regression Analysis
4. Simulation Modeling
5. Lean and Six Sigma methodologies
6. Optimization algorithms (e.g., linear programming, genetic algorithms)

Steps for Process Optimization:

1. Define problem and objectives
2. Collect and analyze data
3. Identify opportunities for improvement
4. Develop and evaluate solutions
5. Implement changes
6. Monitor and adjust

Benefits of Process Optimization:

1. Improved process performance
2. Increased efficiency and productivity
3. Reduced costs and waste
4. Enhanced quality and customer satisfaction
5. Competitive advantage

Common applications of Process Optimization:

1. Manufacturing: Production planning and control
2. Supply Chain Management: Logistics and distribution
3. Service Industries: Quality and efficiency improvement
4. Healthcare: Patient flow and resource allocation

5. Finance: Portfolio optimization and risk management

21.2 USES OF SQC

Statistical Quality Control (SQC) has a wide range of applications across various industries, including:

1. **Manufacturing:** Monitoring production processes, controlling quality, and improving efficiency.

2. **Healthcare**: Improving patient outcomes, reducing medical errors, and optimizing clinical processes.

3**. Service** Industries: Enhancing customer satisfaction, improving service quality, and increasing efficiency.

4. **Finance**: Managing risk, optimizing investment portfolios, and improving financial processes.

5. Supply Chain Management: Ensuring quality and reliability of supplies, managing inventory, and optimizing logistics.

6. **Pharmaceuticals:** Ensuring quality and efficacy of drugs, managing clinical trials, and optimizing manufacturing processes.

7. Aerospace: Ensuring safety and reliability of aircraft and spacecraft, managing complex systems, and optimizing maintenance processes.

8. Automotive: Improving vehicle quality, reducing defects, and optimizing manufacturing processes.

9. Food Processing: Ensuring food safety, managing quality, and optimizing production processes.

10. Software Development: Improving software quality, reducing bugs, and optimizing development processes.

11. Construction: Ensuring quality and safety of buildings, managing construction processes, and optimizing resource allocation.

12. Telecommunications: Ensuring network quality, managing capacity, and optimizing service delivery.

13. Government: Improving public services, managing resources, and optimizing processes.

14. Education: Improving student outcomes, managing resources, and optimizing educational processes.

15. Research and Development: Ensuring quality and validity of research, managing data, and optimizing experimental processes.

21.3 BENEFITS OF SQC

1. Improved quality: Reduced defects and variability.

2. Increased efficiency: Reduced waste and improved productivity.

3. Cost savings: Reduced costs associated with defects and rework.

4. Data-driven decision making: Objective decision making based on statistical analysis.

5. Process optimization: Identification of optimal process settings.

6. Reduced variability: Reduced process variability and improved stability.

7. Improved customer satisfaction: Higher quality products and services.

8. Competitive advantage: Differentiation through high-quality products and services.

9. Reduced scrap and rework: Minimized waste and rework.

10. Improved supply chain management: Better quality and delivery performance.

11. Enhanced employee involvement: Empowered employees through data-driven decision making.

12. Continuous improvement: Culture of ongoing improvement and learning.

13. Reduced testing and inspection: Optimized testing and inspection procedures.

14. Improved product reliability: Increased product reliability and durability.

15. Reduced returns and complaints: Minimized returns and customer complaints.

16. Improved brand reputation: Enhanced reputation through high-quality products and services.

17. Compliance with regulations: Demonstrated compliance with regulatory requirements.

18. Reduced risk: Mitigated risk through proactive quality control.

19. Improved collaboration: Enhanced collaboration across departments and teams.

20. Data-driven innovation: Identification of opportunities for innovation and improvement.

21.4 LIMITATIONS OF STATISTICAL QUALITY CONTROL

Statistical Quality Control (SQC) has several limitations:

1. Assumes normality: Many SQC methods assume normality of data, which may not always be the case.

2. Sensitive to outliers: SQC methods can be affected by outliers, which can lead to incorrect conclusions.

3. Requires large samples: Some SQC methods require large sample sizes to be effective.

4. Does not address root causes: SQC primarily focuses on detecting deviations, not identifying root causes.

5. Can be time-consuming: Collecting and analyzing data can be time-consuming.

6. Requires expertise: SQC requires statistical knowledge and expertise.

7. Not suitable for all processes: SQC may not be applicable to all processes, especially those with complex relationships.

8. Does not account for human factors: SQC focuses on processes, not human factors that can impact quality.

9. Can lead to false sense of security: Relying solely on SQC can lead to complacency.

10. Requires continuous monitoring: SQC requires ongoing monitoring to ensure effectiveness.

11. May not detect non-random variations: SQC may not detect non-random variations or special causes.

12. Can be affected by measurement errors: SQC is only as good as the data, which can be affected by measurement errors.

21.3 FUNCTIONS OF STATISTICAL QUALITY CONTROL

Statistical Quality Control (SQC) functions include:

1. Quality Planning: Defining quality objectives, policies, and procedures.

2. Quality Control: Monitoring and controlling processes to ensure quality.

3. Quality Assurance: Ensuring compliance with quality standards and procedures.

4. Quality Improvement: Identifying and implementing opportunities for improvement.

5. Process Control: Monitoring and controlling processes to ensure stability and capability.

6. Product Control: Monitoring and controlling products to ensure quality and reliability.

7. Inventory Control: Managing inventory to ensure quality and availability.

8. Supply Chain Control: Managing suppliers and supply chain to ensure quality and reliability.

9. Quality Auditing: Conducting audits to ensure compliance with quality standards and procedures.

10. Training and Development: Providing training and development to ensure personnel are competent in SQC methods.

11. Continuous Improvement: Encouraging a culture of continuous improvement and learning.

12. Data Analysis: Analyzing data to identify trends, patterns, and opportunities for improvement.

13. Reporting and Documentation: Maintaining accurate records and reports to ensure transparency and traceability.

14. Corrective Action: Identifying and implementing corrective actions to address quality issues.

15. Preventive Action: Identifying and implementing preventive actions to prevent quality issues.

12.4 SUMMARY

Statistical Quality Control (SQC) is a methodology used to monitor and control processes to ensure they operate within predetermined limits. It employs statistical methods to detect and prevent defects, reducing waste and improving overall quality. Control Charts are a fundamental tool in Statistical Quality Control (SQC). They're graphical representations of process data over time. Process Capability Analysis (PCA) is a statistical method used to evaluate a process's ability to produce output within specified limits or tolerances. Hypothesis Testing is a statistical method used to make inferences about a population based on a sample of data. Design of Experiments (DOE) is a systematic approach to planning, conducting, and analyzing experiments. Process Optimization is the application of methods and tools to improve the efficiency, effectiveness, and quality of processes.

21.5 TEST YOUR KNOWLEDGE

1. What is Statistical quality control?

   .....
   .....
   .....
   .....

2. What are the benefits of SQC?

   .....
   .....
   .....
   .....

3. Explain the functions of SQC?

   .....
   .....
   .....
   .....

4. What are the uses of SQC?

   .....
   .....
   .....
   .....

5. What are the limitations of SQC?

   .....
   .....
   .....

21.6 FURTHER READINGS

1. Sankalp Gaurav, Business Statistics, Agra Book International
2. Alhance D.N. and Alhance Beena, Fundamental of Statistics.
3. Monga, G.S., Elementary Statistics.
4. Gupta, S.B., Principles of Statistics.

UNIT 22 CONSTRUCTION OF CONTROL CHARTS

UNIT STRUCTURE

22.0 OBJECTIVES

- After going through this unit you will be able to have knowledge about
- The construction of control chart
- Different types of control chart
- Control chart constant and formula to calculate control chart

22.1 INTRODUCTION TO CONTROL CHARTS

A control chart is a graphical representation of data that helps to monitor, control, and improve processes. It's a powerful tool used in Statistical Quality Control (SQC) to:

1. Detect deviations: Identify unusual patterns or trends in the data.

2. Determine stability: Check if the process is stable and in control.

3. Improve quality: Reduce variability and improve process performance.

4. Optimize processes: Identify opportunities for process improvement.

22.2 KEY COMPONENTS OF A CONTROL CHART

The key components of a control chart are:

1. Center Line (CL): The average value of the process, representing the expected outcome.
2. Upper Control Limit (UCL): The maximum acceptable value, above which the process is considered out of control.
3. Lower Control Limit (LCL): The minimum acceptable value, below which the process is considered out of control.

4. Data Points: Individual measurements plotted over time, representing the actual process outcomes.
5. Control Limits: The UCL and LCL, which define the acceptable range of variation.
6. Zone Lines: Optional lines dividing the chart into zones, helping to identify trends and patterns.
7. Chart Title: A brief description of the process being monitored.
8. X-axis: The horizontal axis, representing time or sample sequence.
9. Y-axis: The vertical axis, representing the measured value or characteristic.

22.3 HOW CONTROL CHARTS WORK

1. Data Collection: Collect data on the process characteristic being monitored.
2. Plotting: Plot the data points on the control chart over time.
3. Centre Line (CL): Calculate the average value of the process (CL).
4. Control Limits: Calculate the Upper Control Limit (UCL) and Lower Control Limit (LCL) based on the process variability.
5. Interpretation:
    a. In-Control: Data points fluctuate randomly within the control limits, indicating a stable process.
    b. Out-of-Control: Data points exceed the control limits or show non-random patterns, indicating a process issue.
    c. Trends: Data points gradually shift upward or downward, indicating a process drift.
    d. Patterns: Data points exhibit cyclical or systematic behavior, indicating a process anomaly.
6. Investigation: When the process is out-of-control, investigate the cause and take corrective action.
7. Adjustments: Adjust the process to bring it back into control, if necessary.
8. Ongoing Monitoring: Continuously monitor the process to ensure it remains in control.

22.4 TYPES OF CONTROL CHARTS

There are several types of control charts, each designed to monitor specific types of data and processes:

1. $\bar{X}$-R Chart: Monitors process mean ($\bar{X}$) and range (R) for continuous data.
2. $\bar{X}$-s Chart: Monitors process mean ($\bar{X}$) and standard deviation (s) for continuous data.
3. p-Chart: Monitors proportion of defective units (p) for attribute data.
4. c-Chart: Monitors count of defects per unit (c) for attribute data.
5. Individuals-Moving Range (I-MR) Chart: Monitors individual data points and moving range

for continuous data.

6. Cumulative Sum (CUSUM) Chart: Monitors cumulative sum of deviations from the target value.
7. Exponential Weighted Moving Average (EWMA) Chart: Monitors weighted average of past data points.
8. Short-Run SPC Chart: Monitors processes with short production runs or frequent changes.
9. Modified Control Limits Chart: Adjusts control limits based on process changes or improvements.
10. Time-Weighted Control Chart: Gives more weight to recent data points.
11. Multivariate Control Chart: Monitors multiple variables simultaneously.
12. Hotelling's T-Square Chart: Monitors multiple variables and detects outliers.

22.5 CONSTRUCTION OF CONTROL CHARTS

Construction of Control Charts involves the following steps:

1. Define the process: Identify the process to be controlled and the quality characteristic to be measured.
2. Determine the sampling method: Choose a suitable sampling method (e.g., random, stratified).
3. Select the sample size: Determine the sample size (n) based on the process variability and desired level of precision.
4. Collect data: Collect data on the quality characteristic for each sample.
5. Calculate the control limits: Calculate the upper control limit (UCL), lower control limit (LCL), and center line (CL) using statistical formulas.
6. Plot the data: Plot the data points on the control chart.
7. Interpret the chart: Look for trends, patterns, or points outside the control limits to determine if the process is in control.

22.6 FORMULA TO CALCULATE CONTROL CHART

$\bar{X}$-R Chart:

1. Center Line (CL): $\bar{X} = (\Sigma x)/n$

2. Upper Control Limit (UCL): $UCL = \bar{X} + (A2 * R)$

3. Lower Control Limit (LCL): $LCL = \bar{X} - (A2 * R)$

4. Range (R): $R = x(max) - x(min)$

$\bar{X}$-s Chart:

1. Center Line (CL): $\bar{X} = (\Sigma x)/n$

2. Upper Control Limit (UCL): $UCL = \bar{X} + (3 * s/c4)$

3. Lower Control Limit (LCL): LCL = $\bar{X}$ - (3 * s/c4)

4. Standard Deviation (s): s = √[(Σ(x - $\bar{X}$)^2)/(n - 1)]

p-Chart:

1. Center Line (CL): p = (Σx)/n

2. Upper Control Limit (UCL): UCL = p + 3√(p(1-p)/n)

3. Lower Control Limit (LCL): LCL = p - 3√(p(1-p)/n)

c-Chart:

1. Center Line (CL): c = (Σx)/n

2. Upper Control Limit (UCL): UCL = c + 3√(c)

3. Lower Control Limit (LCL): LCL = c - 3√(c)

Individuals-Moving Range (I-MR) Chart:

1. Center Line (CL): $\bar{X}$ = (Σx)/n

2. Upper Control Limit (UCL): UCL = $\bar{X}$ + 3σ

3. Lower Control Limit (LCL): LCL = $\bar{X}$ - 3σ

4. Moving Range (MR): MR = |x(i) - x(i-1)|

22.7 CONTROL CHART CONSTANTS:

Control Chart Constants are used to calculate control limits and are based on the distribution of the data. Here are some common control chart constants:

1. d2: Factor for calculating control limits for $\bar{X}$-R charts (d2 = 1.128 for n=2, 1.693 for n=3, etc.)

2. d3: Factor for calculating control limits for $\bar{X}$-R charts (d3 = 0 for n=2, 1.772 for n=3, etc.)

3. A2: Factor for calculating control limits for $\bar{X}$-R charts (A2 = 0.73 for n=2, 0.577 for n=3, etc.)

4. D3: Factor for calculating control limits for $\bar{X}$-s charts (D3 = 0 for n=2, 1.128 for n=3, etc.)

5. D4: Factor for calculating control limits for $\bar{X}$-s charts (D4 = 3.267 for n=2, 2.574 for n=3, etc.)

6. c4: Factor for calculating control limits for individuals-Moving Range (I-MR) charts (c4 = 0.7979 for n=2, 0.8862 for n=3, etc.)

These constants are used in the formulas to calculate the control limits for different types of control charts. The values of these constants depend on the sample size (n) and the type of chart being used.

22.8 SUMMARY

A control chart is a graphical representation of data that helps to monitor, control, and improve processes. It's a powerful tool used in Statistical Quality Control (SQC), Centre Line (CL), Upper Control Limit (UCL), Lower Control Limit (LCL), Data Points, Control Limits, Zone Lines, Chart Title, X-axis& Y-axis are the key components of control chart.

22.9 TEST YOUR KNOWLEDGE

1. What are the steps to construct control chart
2. What is control chart? What are its key components?
3. Explain the key components of control chart?
4. What is the control chart constant? Explain
5. Write formula of any three control chart?

22.10 FURTHER READINGS

1. Sankalp Gaurav, Business Statistics, Agra Book International
2. Alhance D.N. and Alhance Beena, Fundamental of Statistics.
3. Monga, G.S., Elementary Statistics.
4. Gupta, S.B., Principles of Statistics.

Unit 23 TIME SERIES

UNIT STRUCTURE

OBJECTIVES

After going through this unit you should be able to know about the–

About Time Series

- Relevance and uses of Time series
- Components of Time series
- Mathematical models for Time series analysis

23.1 INTRODUCTION

A time series is a sequence of data points that occur in successive order over some period of time. This can be contrasted with cross-sectional data, which captures a point-in-time.

In investing, a time series tracks the movement of the chosen data points, such as a security's price, over a specified period of time with data points recorded at regular intervals. There is no minimum or maximum amount of time that must be included, allowing the data to be gathered in a way that provides the information being sought by the investor or analyst examining the activity.

23.2 MEANING OF TIME SERIES

A time series can be taken on any variable that changes over time. In investing, it is common to use a time series to track the price of a security over time. This can be tracked over the short term, such as the price of a security on the hour over the course of a business day, or the long term, such as the price of a security at close on the last day of every month over the course of five years.

Time series analysis can be useful to see how a given asset, security, or economic variable changes over time. It can also be used to examine how the changes associated with the chosen data point compare to shifts in other variables over the same time period.

23.3 IMPORTANCE OF TIME SERIES:

1. Analysis of causes and conditions prevailing during occurrence of past changes, one can easily

determine the future policies and programs

2. Estimation of future trends on the basis of analysis or past trends.

3. Trends of trade cycles are studied and their effect can be reduced to a considerable extent.

4. Comparative study with the other time series.

A time series depicts the relationship between two variables. Time is one of those variables and the second is any quantitative variable. It is not necessary that the relationship always shows increment in the change of the variable with reference to time. The relation is not always decreasing too.

It may be increasing for some and decreasing for some points in time. Can you think of any such example? The temperature of a particular city in a particular week or a month is one of those examples.

23.4 USES OF TIME SERIES

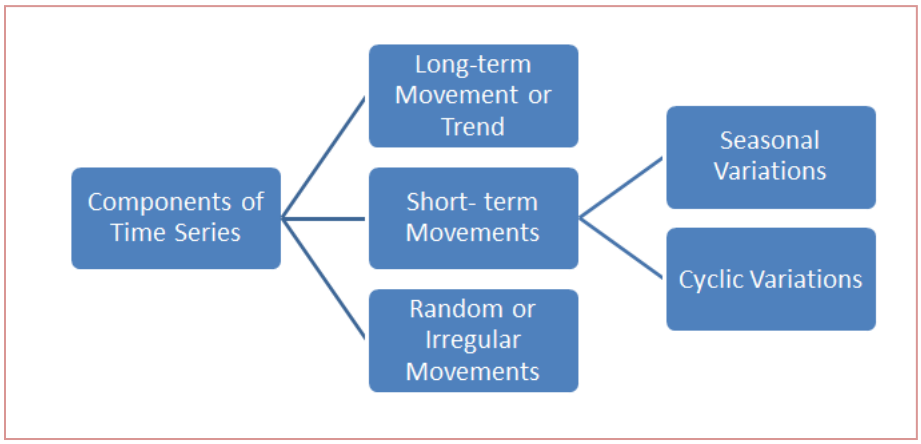
- The most important use of studying time series is that it helps us to predict the future behaviour of the variable based on past experience
- It is helpful for business planning as it helps in comparing the actual current performance with the expected one
- From time series, we get to study the past behaviour of the phenomenon or the variable under consideration
- We can compare the changes in the values of different variables at different times or places, etc.

23.5 Components for Time Series Analysis

The various reasons or the forces which affect the values of an observation in a time series are the components of a time series. The four categories of the components of time series are

- Trend
- Seasonal Variations
- Cyclic Variations
- Random or Irregular movements

Seasonal and Cyclic Variations are the periodic changes or short-term fluctuations.



Trend

The trend shows the general tendency of the data to increase or decrease during a long period of time. A trend is a smooth, general, long-term, average tendency. It is not always necessary that the increase or decrease is in the same direction throughout the given period of time.

It is observable that the tendencies may increase, decrease or are stable in different sections of time. But the overall trend must be upward, downward or stable. The population, agricultural production, items manufactured, number of births and deaths, number of industry or any factory, number of schools or colleges are some of its example showing some kind of tendencies of movement.

Linear and Non-Linear Trend

If we plot the time series values on a graph in accordance with time t. The pattern of the data clustering shows the type of trend. If the set of data cluster more or less round a straight line, then the trend is linear otherwise it is non-linear (Curvilinear).

Periodic Fluctuations

There are some components in a time series which tend to repeat themselves over a certain period of time. They act in a regular spasmodic manner.

Seasonal Variations

These are the rhythmic forces which operate in a regular and periodic manner over a span of less than a year. They have the same or almost the same pattern during a period of 12 months. This variation will be present in a time series if the data are recorded hourly, daily, weekly, quarterly, or monthly.

These variations come into play either because of the natural forces or man-made conventions. The various seasons or climatic conditions play an important role in seasonal variations. Such as production of crops depends on seasons, the sale of umbrella and raincoats in the rainy season, and the sale of electric fans and A.C. shoots up in summer seasons.

The effect of man-made conventions such as some festivals, customs, habits, fashions, and some occasions like marriage is easily noticeable. They recur themselves year after year. An upswing in a season should not be taken as an indicator of better business conditions.

Cyclic Variations

The variations in a time series which operate themselves over a span of more than one year are the cyclic variations. This oscillatory movement has a period of oscillation of more than a year. One complete period is a cycle. This cyclic movement is sometimes called the ‘Business Cycle’.

It is a four-phase cycle comprising of the phases of prosperity, recession, depression, and recovery. The cyclic variation may be regular are not periodic. The upswings and the downswings in business depend upon the joint nature of the economic forces and the interaction between them.

Random or Irregular Movements

There is another factor which causes the variation in the variable under study. They are not regular variations and are purely random or irregular. These fluctuations are unforeseen, uncontrollable, unpredictable, and are erratic. These forces are earthquakes, wars, flood, famines, and any other disasters.

23.6 MATHEMATICAL MODEL FOR TIME SERIES ANALYSIS

Mathematically, a time series is given as

$y_t = f(t)$

Here, y_t is the value of the variable under study at time t. If the population is the variable under study at the various time period $t_1, t_2, t_3, \dots, t_n$. Then the time series is

t: $t_1, t_2, t_3, \dots, t_n$

y_t: $y_{t1}, y_{t2}, y_{t3}, \dots, y_{tn}$

or, t: $t_1, t_2, t_3, \dots, t_n$

y_t: $y_1, y_2, y_3, \dots, y_n$

23.6.1 Additive Model for Time Series Analysis

If y_t is the time series value at time t. T_t, S_t, C_t, and R_t are the trend value, seasonal, cyclic and random fluctuations at time t respectively. According to the Additive Model, a time series can be expressed as

$y_t = T_t + S_t + C_t + R_t.$

This model assumes that all four components of the time series act independently of each other.

23.6.2 Multiplicative Model for Time Series Analysis

The multiplicative model assumes that the various components in a time series operate proportionately to each other. According to this model

$y_t = T_t \times S_t \times C_t \times R_t$

23.6.3 Mixed models

Different assumptions lead to different combinations of additive and multiplicative models as

$y_t = T_t + S_t + C_t R_t.$

The time series analysis can also be done using the model $y_t = T_t + S_t \times C_t \times R_t$ or $y_t = T_t \times C_t + S_t \times R_t$ etc.

23.7 MEASURES OF SECULAR TREND:

1. Freehand Curve Method

2. Smoothing Methods: Semi Average Method

3. Moving Average Method

4. Trend Projection Methods

5. Least Square Method

1. FREEHAND CURVE METHOD: It is the most simplest method according to this method firstly time series is plotted on the graph paper, keeping in view the direction of fluctuation of the time series straight line or curve is drawn passing through the midpoints. The line or curve represents the secular trend.

2. SEMI-AVERAGE METHOD: According to this method the original data is divided into two

parts. Then the mean of both the parts is calculated separately.

3. MOVING AVERAGE METHOD : This is the simple and most widely used method for the calculation of trend values. In this firstly we have to decide what will be the period of moving average? The period can be odd i.e 3, 5, 7, 9, 11 or even i.e 2, 4, 6, 8, 10 years

4. TREND PROJECTION METHODS: It is best represented by straight line is termed as long run directions(upward, downward, constant), of any business activity over a period of several years. Reasons to study trend: Helps in describing long term general directions of any business activity over a long period of time. It helps in making intermediate and long term forecasting projections.

5. LEAST SQUARE METHOD: It is calculated to be the best method for calculating the trend values. The trend line obtained by this methods is called line of best fit. This line can be a straight line or parabolic curve. Least Square Method: This method can be used in both cases when no. of years are odd as well as even. $\sum Y = N a + b \sum X \sum XY = a \sum X + b \sum X 2$ Function Equation $Y = a + b x$

Lets understand two most important methods in detail

- Moving average
- Least square method

Moving average methods

The moving average of a period (extent) *m* is a series of successive averages of *m* terms at a time. The data set used for calculating the average starts with first, second, third and etc. at a time and *m* data taken at a time.

In other words, the first average is the mean of the first *m* terms. The second average is the mean of the *m* terms starting from the second data up to $(m + 1)^{th}$ term. Similarly, the third average is the mean of the *m* terms from the third to $(m + 2)^{th}$ term and so on.

If the extent or the period, *m* is odd i.e., *m* is of the form (2k + 1), the moving average is placed against the mid-value of the time interval it covers, i.e., t = k + 1. On the other hand, if *m* is even i.e., *m* = 2k, it is placed between the two middle values of the time interval it covers, i.e., t = k and t = k + 1.

When the period of the moving average is even, then we need to synchronize the moving average with the original time period. It is done by centring the moving averages i.e., by taking the average of the two successive moving averages.

Advantages of Moving average

Some of the advantages of using moving averages include:

- Moving average is used for forecasting goods or commodities with constant demand, where there is a slight trend or seasonality.
- Moving average is useful for separating out random variations.
- Moving average can help you identify areas of support and resistance.
- Simplicity of application and interpretation makes it possible to plot several different moving average lines at the same time.
- Moving average gives constant forecasts.

- Disadvantages of Moving Average
- The main problem is to determine the extent of the moving average which completely eliminates the oscillatory fluctuations.
- This method assumes that the trend is linear but it is not always the case.
- It does not provide the trend values for all the terms.
- This method cannot be used for forecasting future trend which is the main objective of the time series analysis.

Calculation of Moving averages

Moving Averages Method gives a trend with a fair degree of accuracy. In this method, we take arithmetic mean of the values for a certain time span. The time span can be three-years, four -years, five- years and so on depending on the data set and our interest. We will see the working procedure of this method.

Procedure:

(i) Decide the period of moving averages (three- years, four -years).

(ii) In case of odd years, averages can be obtained by calculating,

$$\frac{a+b+c}{3}, \frac{b+c+d}{3}, \frac{c+d+e}{3}, \frac{d+e+f}{3}, \ldots\ldots$$

(iii) If the moving average is an odd number, there is no problem of centering it, the average value will be centered besides the second year for every three years.

(iv) In case of even years, averages can be obtained by calculating,

$$\frac{a+b+c+d}{4}, \frac{b+c+d+e}{4}, \frac{c+d+e+f}{4}, \frac{d+e+f+g}{4}, \ldots$$

(v) If the moving average is an even number, the average of first four values will be placed between 2 nd and 3rd year, similarly the average of the second four values will be placed between 3rd and 4th year. These two averages will be again averaged and placed in the 3rd year. This continues for rest of the values in the problem. This process is called as centering of the averages.

Example of time moving average method (For Odd Years)

Give 3 year moving average of the following series

Year	Production (in lakhs Tons)
2014	15
2015	15.2

2016	16
2017	16.3
2018	17
2019	17.5
2020	17.6

Solution

Year	Production (in lakhs Tons)	3 Years Total	3 Years Moving Average , Trends
2014	15	.....	
2015	15.2	46.2	15.4
2016	16	47.5	15.8
2017	16.3	49.3	16.4
2018	17	50.8	16.9
2019	17.5	52.1	17.3
2020	17.6	.....	.....

Example of time moving average method (For Even Years)

Give 4 year moving average of the following series

Year	Production (in lakhs Tons)
2014	15
2015	15.2
2016	16
2017	16.3
2018	17

2019	17.5
2020	17.6

Solution

Year	Production (in lakhs Tons)	4 Years Total	2 year total of 4 year total	4Years Moving Average , Trends
2014	15	.....		
		.		
2015	15.2			
		62.5		
2016	16		127	15.87
		64.5		
2017	16.3		131.3	16.41
		66.8		
2018	17		135.2	16.9
		68.4		
2019	17.5	.....		
2020	17.6	.....		.....

Note: for taking the 4 year moving average the value of 2 year total is divided from the double of asked year value i.e in this question they are asking 4 years so the value is divided from 8)

Method of Least Squares

During Time Series analysis we come across with variables, many of them are dependent upon others. It is often required to find a relationship between two or more variables. Least Square is the method for finding the best fit of a set of data points. It minimizes the sum of the residuals of points from the plotted curve. It gives the trend line of best fit to a time series data. This method is most widely used in time series analysis. Let us discuss the Method of Least Squares in detail.

Each point on the fitted curve represents the relationship between a known independent variable and an unknown dependent variable.

In general, the least squares method uses a straight line in order to fit through the given points which are known as the method of linear or ordinary least squares. This line is termed as the line of best fit from which the sum of squares of the distances from the points is minimized.

Equations with certain parameters usually represent the results in this method. The method of least squares actually defines the solution for the minimization of the sum of squares of deviations or the errors in the result of each equation.

The least squares method is used mostly for data fitting. The best fit result minimizes the sum of squared errors or residuals which are said to be the differences between the observed or experimental value and corresponding fitted value given in the model. There are two basic kinds of the least squares methods – ordinary or linear least squares and nonlinear least squares.

ADVANTAGES OF LEAST SQUARE METHOD

(i) This method is completely free from personal bias of the analyst as it is very objective in nature. Any body using this method is bound to fit the same type of straight line, and find the same trend values for the series.

(ii) Unlike the moving average method, under this method, we are able to find the trend values for the entire time series without any exception for the extreme periods of the series even.

(iii) Unlike the moving average method, under this method, it is quite possible to forecast any past of future values perfectly, since the method provides us with a functional relationship between two variables in the form of a trend line equation, viz. $Yc = a + bX$, $Yc = a + bX + cX^2 + \ldots$ Or $Yc = ab^X$ etc.

(iv) This method provides us with a rate of growth per period i.e. b, as shown in the equations cited above. With this rate of growth, we can very well determine the value for any past or previous year by the process of successive addition, or deduction from the trend values of the origin of X.

(v) This method providers us with the line of the best fit from which the sum of the positive and negative deviation is zero, and the sum of the squares of the deviations is least i.e.

(i) $\sum (Y-Yc) = 0$; and (ii)) $\sum (Y-Yc)^2$ = The least values.

(vi) This method is the most popular, and widely used for fitting mathematical function to a given set of observations.

This method is very flexible in the sense that it allows for shifting the trend origin from one point of time to another, and for the conversion of the annual trend equation into monthly, or quarterly trend trend equation, and vice versa.

DEMERITS OF LEAST SQUARE METHOD

(i) This method is very much rigid in the sense that if any item is added to, or subtracted from the series, it will need a thorough revision of the trend equation to fit a trend line, and find the trend values thereby.

(ii) In comparison to the other methods of trend determination, the method is bit complicated in as much as it involves many mathematical tabulations, computations, and solutions like those of simultaneous equations.

(iii) Under this method, we forecast the past and future values basing upon the trend values only, and we do not take note of the seasonal, cyclical and irregular components of the series for the purpose.

(iv) This method is not suitable for business, and economic data which conform to the growth curves like Gompertz's curve, Logistic Pearl-Read curve etc.

(v) It needs great care for the determination of the type of the trend curve to be fitted in viz : linear, parabolic, exponential, or any other more complicated curve. An erratic selection of the type of curve may lead to fallacious conclusions.

(vi) This method is quite inappropriate for both very short and very long series. It is also unsuitable for a series in which the differences between the successive observations are not found to be constant, or nearly so.

Mathematical Representation

It is a mathematical method and with it gives a fitted trend line for the set of data in such a manner that the following two conditions are satisfied.

1. The sum of the deviations of the actual values of Y and the computed values of Y is zero.
2. The sum of the squares of the deviations of the actual values and the computed values is least.

This method gives the line which is the line of best fit. This method is applicable to give results either to fit a straight line trend or a parabolic trend.

The method of least squares as studied in time series analysis is used to find the trend line of best fit to a time series data.

Secular Trend Line

The secular trend line (Y) is defined by the following equation:

$Y = a + b X$

Where, Y = predicted value of the dependent variable

a = Y-axis intercept i.e. the height of the line above origin (when X = 0, Y = a)

b = slope of the line (the rate of change in Y for a given change in X)

When b is positive the slope is upwards, when b is negative, the slope is downwards

X = independent variable (in this case it is time)

To estimate the constants a and b, the following two equations have to be solved simultaneously:

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

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

To simplify the calculations, if the midpoint of the time series is taken as origin, then the negative values in the first half of the series balance out the positive values in the second half so that $\Sigma X = 0$. In this case, the above two normal equations will be as follows:

$\Sigma Y = na$

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

In such a case the values of a and b can be calculated as under:

Since $\Sigma Y = na$

$a = \sum Yn$

Since, $\Sigma XY = b\Sigma X^2$

Solved Example on Method of Least Squares

Fit a straight line trend on the following data using the Least Squares Method.

Period (year)	2014	2015	2016	2017	2018	2019	2020	2021	2022
Y	4	7	7	8	9	11	13	14	17

Solution: Total of 9 observations are there. So, the origin is taken at the Year 2000 for which X is assumed to be 0.

PERIOD (YEAR)	**Y**	**X**	**XY**	**X^2**	**REMARK**
2014	4	-4	-16	16	
2015	7	-3	-21	9	
2016	7	-2	-14	4	NEGATIVE REGION
2017	8	-1	-8	1	
2017	9	0	0	0	ORIGIN
2019	11	1	11	1	
2020	13	2	16	4	POSITIVE REGION

2021	14	3	42	9
2022	17	4	68	16
Total (Σ)	ΣY = 90	ΣX = 0	ΣXY = 88	SΣX^2 =60

From the table we find that value of n is 9, value of ΣY is 90, value of ΣX is 0, value of ΣXY is 88 and value of ΣX^2 is 60 .

Substituting these values in the two given equations,

a = 909 or a = 10
b = 8860 or b = 1.47
Trend equation is : Y = 10 + 1.47 X

Question Calculate the seasonal indices from the following data using the average from the following data using the average method:

	I Quarterly	II Quarterly	III Quarterly	IV Quarterly
2008	72	68	62	76
2009	78	74	78	72
2010	74	70	72	76
2011	76	74	74	72
2012	72	72	76	68

Answers:

	I	II	III	IV
Total	372	358	362	364
Average	74.4	71.6	72.4	72.8
Seasonal indices	102.19	98.35	99.45	100

Grand Average = 72.8

Question: Compute 4-year moving averages centered for the following time series:

Years	1995	1996	1997	1998	1999	2000	2001	2002

Production	80	90	92	83	87	96	100	110

Solution:

Year	Production	4-Year Moving Total	4-Year Moving Average	2-values Moving Total	4-year Moving Average Centered
1995	80	—	—	—	—
1996	90	345	86.25	—	—
1997	92	352	88.00	174.25	87.125
1998	83	358	89.50	177.50	88.750
1999	87	366	91.50	181.00	90.500
2000	96	393	98.25	189.75	94.875
2001	100	—	—	—	—
2002	110	—	—	—	—

Question: Compute 5-year, 7-year and 9-year moving averages for the following data.

Years	1990	1991	1992	1993	1994	1995	1996	19997	1998	1999	2000
Values	2	4	6	8	10	12	14	16	18	20	22

Solution:

The necessary calculations are given below:

		5-Year Moving		7-Year Moving		9-Year Moving	
Years	**Values**	**Total**	**Average**	**Total**	**Average**	**Total**	**Average**
1990	2	—	—	—	—	—	—

1991	4	—	—	—	—	—	—
1992	6	30	6	—	—	—	—
1993	8	40	8	56	8	—	—
1994	10	50	10	70	10	90	10
1995	12	60	12	84	12	108	12
1996	14	70	14	98	14	126	14
1997	16	80	16	112	16	—	
1998	18	90	18	—	—	—	—
1999	20	—	—	—	—	—	—
2000	22	—	—	—	—	—	—

23.8 TEST YOUR KNOWLEDGE

1. What is time Series
2. Explain trend
3. Explain seasonal fluctuations
4. Define least square method
5. What is moving average method?
6. With imaginary figure construct a 7 year moving average?
7. Explain cyclic variations
8. Discuss about irregular variation
9. Define seasonal index.
10. Explain the method of fitting a straight line.
11. State the two normal equations used in fitting a straight line.
12. State the different methods of measuring trend.

23.9 FURTHER READINGS

1. Sankalp Gaurav, Business Statistics, Agra Book International

2. Alhance D.N. and Alhance Beena, Fundamental of Statistics.
3. Monga, G.S., Elementary Statistics.
4. Gupta, S.B., Principles of Statistics.

UNIT 24 CHI SQUARE TEST

UNIT STRUCTURE

Objectives

After going through this unit you should be able to know about the–

i. Conceptual Framework of Chi Square Test

ii. Test of goodness of fit

24.1 INTRODUCTION

The chi-square test is an important test amongst the several tests of significance developed by statisticians. Chi-square, symbolically written as χ^2 (Pronounced as Ki-square), is a statistical measure used in the context of sampling analysis for comparing a variance to a theoretical variance. As a non-parametric* test, it "can be used to determine if categorical data shows dependency or the two classifications are independent. It can also be used to make comparisons between theoretical populations and actual data when categories are used." Thus, the chi-square test is applicable in large number of problems. The test is, in fact, a technique through the use of which it is possible for all researchers to (i) test the goodness of fit; (ii) test the significance of association between two attributes, and (iii) test the homogeneity or the significance of population variance.

Chi-Square as a test for comparing variance

The chi-square value is often used to judge the significance of population variance i.e., we can use the test to judge if a random sample has been drawn from a normal population with mean (□) and with a specified variance (σ_p^2). The test is based on χ^2 – distribution. Such a distribution we encounter when we deal with collections of values that involve adding up squares. Variances of samples require us to add a collection of squared quantities and, thus, have distributions that are related to χ^2 – distribution. If we take each one of a collection of sample variances, divided them by the known population variance and multiply these quotients by (n – 1), where n means the number of items in the sample, we shall obtain a χ^2 – distribution. Thus, $\frac{\sigma_s^2}{\sigma_p^2}(n-1) = \frac{\sigma_s^2}{\sigma_P^2}$ (d.f.) would have the same distribution as χ^2 –distribution with (n –1) **DEGREES OF FREEDOM**

The χ^2 –distribution is not symmetrical and all the values are positive. For making use of this distribution, one is required to know the degrees of freedom since for different degrees of freedom we have different curves. The smaller the number of degrees of freedom, the more skewed is the distribution

In brief, when we have to use chi-square as a test of population variance, we have to work out the value of χ^2 to test the null hypothesis (viz., $H_0 : \sigma_s^2 = \sigma_p^2$) as under :

$$\chi^2 = \frac{\sigma_s^2}{\sigma_p^2}(n-1)$$

where

σ_s^2 = variance of the sample;

σ_p^2 = variance of the population;

(n–1) = degrees of freedom, n being the number of items in the sample.

Then by comparing the calculated value with the table value of χ^2 for (n – 1) degrees of freedom at a given level of significance, we may either accept or reject the null hypothesis. If the calculated value of χ^2 is less than the table value, the null hypothesis is accepted, but if the calculated value is equal or greater than the table value, the hypothesis is rejected. All this can be made clear by an example.

Illustration 1

Weight of 10 students is as follows :

S.No.	1	2	3	4	5	6	7	8	9	10
Weight (kg.)	38	40	45	53	47	43	55	48	52	49

Can we say that the variance of the distribution of weight of all students from which the above sample of 10 students was drawn is equal to 20 kgs? Test this at 5 per cent and 1 per cent level of significance.

Solution

First of all we should work out the variance of the sample data or σ_s^2 and the same has been worked out as under:

S.No.	**X_i (Weight in kgs.)**	$(X_i - \bar{X})$	$(X_i - \bar{X})^2$
1	38	–9	81
2	40	–7	49
3	45	–2	04
4	53	+6	36
5	47	+0	00
6	43	–4	16
7	55	+8	64
8	48	+1	01
9	52	+5	25
10	49	+2	04
n = 10	**$\sum X_i = 470$**		**$\sum (X_i - \bar{X})^2 = 280$**

$$\bar{X} = \frac{\sum X_i}{n} = \frac{470}{10} = 47 \text{ kgs.}$$

$$\therefore \quad \sigma_s = \sqrt{\frac{\sum (X_i - \bar{X})^2}{n-1}} = \sqrt{\frac{280}{10-1}} = \sqrt{31.11}$$

or $$\sigma_s^2 = 31.11.$$

Let the null hypothesis be $H_0 : \sigma_p^2 = \sigma_s^2$. In order to test this hypothesis we work out the χ^2 value as under:

$$\chi^2 = \frac{\sigma_s^2}{\sigma_p^2}(n-1)$$

$$= \frac{31.11}{20}(10-1) = 13.999$$

Degrees of freedom in the given case is (n – 1) = (10 – 1) = 9. At 5 per cent level of significance the table value of χ^2 = 16.92 and at 1 per cent level of significance, it is 21.67 for 9 d.f. and both these values are greater than the calculated value of χ^2 which is 13.999. Hence we accept the null hypothesis and conclude that the variance of the given distribution can be taken as 20 kgs at 5 per cent as also at 1 per cent level of significance. In other words, the sample can be said to have been taken from a population with variance 20 kgs .

Illustration 2

A sample of 10 is drawn randomly from a certain population. The sum of the squared deviations from the mean of the given sample is 50. Test the hypothesis that the variance of the population is 5 at 5 per cent level of significance.

Solution.

n = 10

$$\sum (X_i - \bar{X})^2 = 50$$

$$\sigma_s^2 = \frac{\sum (X_i - \bar{X})^2}{n-1} = \frac{50}{9}$$

Take the null hypothesis as $H_0 : \sigma_p^2 = \sigma_s^2$. In order to test this hypothesis, we work out the χ^2 value as under:

$$\chi^2 = \frac{\sigma_s^2}{\sigma_p^2}(n-1) = \frac{\frac{50}{9}}{5}(10-1) = \frac{50}{9} \times \frac{1}{5} \times \frac{9}{1} = 10$$

Degrees of freedom = (10 – 1) = 9.

The table value of χ^2 at 5 per cent level for 9 d.f. is 16.92. The calculated value of χ^2 is less than this table value, so we accept the null hypothesis and conclude that the variance of the population is 5 as given in the question.

24.2 CHI-SQUARE AS A NON-PARAMETRIC TEST

Chi-square is an important non-parametric test and as such no rigid assumptions are necessary in respect of the type of population. We require only the degrees of freedom (implicitly of course the size of the sample) for using this test. As a non-parametric test, chi-square can be used (i) as a test of goodness of fit and (ii) as a test of independence.

As a test of goodness of fit, χ^2 test enables us to see how well does the assumed theoretical distribution (such as Binomial distribution, Poisson distribution or Normal distribution) fit to the observed data. When some theoretical distribution is fitted to the given data, we are always interested in knowing as to how well this distribution fits with the observed data. The chi-square test can give answer to this. If the calculated value of χ^2 is less than the table value at a certain level of significance, the fit is considered to be a good one which means that the divergence between the observed and expected frequencies is attributable to fluctuations of sampling. But if the calculated value of χ^2 is greater than its table value, the fit is not considered to be a good one.

As a test of independence, χ^2 test enables us to explain whether or not two attributes are associated. For instance, we may be interested in knowing whether a new medicine is effective in controlling fever or not, χ^2 test will helps us in deciding this issue. In such a situation, we proceed with the null hypothesis that the two attributes (viz., new medicine and control of fever) are independent which means that new medicine is not effective in controlling fever. On this basis we first calculate the expected frequencies and then work out the value of χ^2. If the calculated value of χ^2 is less than the table value at a certain level of significance for given degrees of freedom, we conclude that null hypothesis stands which means that the two attributes are independent or not associated (i.e., the new medicine is not effective in controlling the fever). But if the calculated value of χ^2 is greater than its table value, our inference then would be that null hypothesis does not hold good which means the two attributes are associated and the association is not because of some chance factor but it exists in reality (i.e., the new medicine is effective in controlling the fever and as such may be prescribed). It may, however, be stated here that χ^2 is not a measure of the degree of relationship or the form of relationship between two attributes, but is simply a technique of judging the significance of such association or relationship between two attributes.

In order that we may apply the chi-square test either as a test of goodness of fit or as a test to judge the significance of association between attributes, it is necessary that the observed as well as theoretical or expected frequencies must be grouped in the same way and the theoretical distribution must be adjusted to give the same total frequency as we find in case of observed distribution. χ^2 is then calculated as follows :

$$\chi^2 = \sum \frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}}$$

where,

O_{ij} = observed frequency of the cell in ith row and jth column.

Eij = expected frequency of the cell in ith row and jth column.

If two distributions (observed and theoretical) are exactly alike, $\chi^2 = 0$; but generally due to sampling errors, χ^2 is not equal to zero and as such we must know the sampling distribution of χ^2 so that we may find the probability of an observed χ^2 being given by a random sample from the hypothetical universe. Instead of working out the probabilities, we can use ready table which gives probabilities for given values of χ^2. Whether or not a calculated value of χ^2 is significant can be ascertained by looking at the tabulated values of χ^2 for given degrees of freedom at a certain level of significance. If the calculated value of χ^2 is equal to or exceeds the table value, the difference between the observed and expected frequencies is taken as significant, but if the table value is more than the calculated value of χ^2, then the difference is considered as insignificant i.e., considered to have arisen as a result of chance and as such can be ignored.

As already stated, degrees of freedom play an important part in using the chi -square distribution and the test based on it, one must correctly determine the degrees of freedom. If there are 10 frequency classes and there is one independent constraint, then there are (10 – 1) = 9 degrees of freedom. Thus, if 'n' is the number of groups and one constraint is placed by making the totals of observed and expected frequencies equal, the d.f. would be equal to (n – 1). In the case of a contingency table (i.e., a table with 2 columns and 2 rows or a table with two columns and more than two rows or a table with two rows but more than two columns or a table with more than two rows and more than two columns), the d.f. is worked out as follows:

$$\text{d.f.} = (c - 1)(r - 1)$$

where 'c' means the number of columns and 'r' means the number of rows.

The following conditions should be satisfied before χ^2 test can be applied:

(i) Observations recorded and used are collected on a random basis.

(ii) All the items in the sample must be independent.

(iii) No group should contain very few items, say less than 10. In case where the frequencies are less than 10, regrouping is done by combining the frequencies of adjoining groups so that the new frequencies become greater than 10. Some statisticians take this number as 5, but 10 is regarded as better by most of the statisticians.

(iv) The overall number of items must also be reasonably large. It should normally be at least 50, howsoever small the number of groups may be.

(v) The constraints must be linear. Constraints which involve linear equations in the cell frequencies of a contingency table (i.e., equations containing no squares or higher powers of the frequencies) are known are know as linear constraints.

24.3 TEST OF GOODNESS OF FIT

The goodness of fit of a statistical model describes how well it fits a set of observations. Measures of goodness of fit typically summarize the discrepancy between observed values and the values expected under the model in question. Such measures can be used in statistical hypothesis testing, ego to test for normality of residuals, to test whether two samples are drawn from identical distributions, or whether outcome frequencies follow a specified distribution. In the analysis of variance, one of the components into which the variance is partitioned may be a lack-of-fit sum of squares.

24.4 SUMMARY

A hypothesis is a proposed explanation for a phenomenon. Hypothesis comes from the Greek, Actually, hypothesis refers to a clever idea. But in recent century, it refers to a provisional idea whose merit requires evaluation.

24.5 TEST YOUR KNOWLEDGE

1 What is chi square test? Define

.....
.....
.....
.....

2. Explain goodness of fit

.....
.....
.....
.....

24.6. FURTHER READINGS

1. Sankalp Gaurav, Business Statistics, Agra Book International
2. Alhance D.N. and Alhance Beena, Fundamental of Statistics.
3. Monga, G.S., Elementary Statistics.
4. Gupta, S.B., Principles of Statistics.