ਜਗਤ ਗੁਰੂ ਨਾਨਕ ਦੇਵ
ਪੰਜਾਬ ਸਟੇਟ ਓਪਨ ਯੂਨੀਵਰਸਿਟੀ
ਪਟਿਆਲਾ

The Motto of Our University
(SEWA)
SKILL ENHANCEMENT
EMPLOYABILITY
WISDOM
ACCESSIBILITY

JAGAT GURU NANAK DEV
PUNJAB STATE OPEN UNIVERSITY, PATIALA

(Established by Act No. 19 of 2019 of the Legislature of State of Punjab)

SELF-INSTRUCTIONAL STUDY MATERIAL FOR JGND PSOU

CERTIFICATE/ DIPLOMA IN STATISTICAL ANALYSIS AND RESEARCH METHODOLOGY

SEMESTER II

SARM 4: TIME SERIES ANALYSIS AND PROBABILITY DISTRIBUTIONS

Head Quarter: C/28, The Lower Mall, Patiala-147001
Website: www.psou.ac.in

The Study Material has been prepared exclusively under the guidance of Jagat Guru Nanak Dev Punjab State Open University, Patiala, as per the syllabi prepared by the Committee of experts and approved by the Academic Council.

COURSE COORDINATOR AND EDITOR:

Dr. Pinky Sra

Assistant Professor

JGND PSOU, Patiala.

ਜਗਤ ਗੁਰੂ ਨਾਨਕ ਦੇਵ
ਪੰਜਾਬ ਸਟੇਟ ਓਪਨ ਯੂਨੀਵਰਸਿਟੀ
ਪਟਿਆਲਾ

\T GURU NANAK DEV PUNJAB STATE OPEN UNIVERSITY,PATIALA
(Established by Act No. 19 of 2019 of the Legislature of State of Punjab)

PREFACE

Jagat Guru Nanak Dev Punjab State Open University, Patiala was established in December 2019 by Act 19 of the Legislature of State of Punjab. It is the first and only Open University of the State, entrusted with the responsibility of making higher education accessible to all, especially to those sections of society who do not have the means, time or opportunity to pursue regular education.

In keeping with the nature of an Open University, this University provides a flexible education system to suit every need. The time given to complete a programme is double the duration of a regular mode programme. Well-designed study material has been prepared in consultation with experts in their respective fields.

The University offers programmes which have been designed to provide relevant, skill-based and employability-enhancing education. The study material provided in this booklet is self- instructional, with self-assessment exercises, and recommendations for further readings. The syllabus has been divided in sections, and provided as units for simplification.

The University has a network of 110 Learner Support Centres/Study Centres, to enable students to make use of reading facilities, and for curriculum-based counseling and practicals. We, at the University, welcome you to be a part of this institution of knowledge.

Prof. G.S. Batra
Dean Academic Affairs

CERTIFICATE/ DIPLOMA IN STATISTICAL ANALYSIS AND RESEARCH METHODOLOGY
SEMESTER II
SARM 4: TIME SERIES ANALYSIS AND PROBABILITY DISTRIBUTIONS

Max. Marks: 100
External: 70
Internal: 30
Pass: 40%
Credits: 6

OBJECTIVES:

Give the knowledge regarding probability theory of outcome of real-life experiments through statistical distributions.

INSTRUCTIONS FOR THE PAPER SETTER/ EXAMINER:

1. The syllabus prescribed should be strictly adhered to.
2. The Question Paper will have 70 Multiple-choice questions (MCQs) and four choices of answers will be there covering the entire syllabus. Each question will carry 1 mark. All questions will be compulsory; hence candidates will attempt all the questions.
3. Paper-setters/Examiners are requested to distribute the questions from Section A and Section B of the syllabus equally i.e., 35 questions from Section A and 35 questions from Section B.
4. The examiner shall give clear instructions to the candidates to attempt questions.
5. The duration of each paper will be two hours.

INSTRUCTIONS FOR THE STUDENTS

The question paper shall consist of 70 Multiple-choice questions. All questions will be compulsory and each question will carry 1 mark. There will be no negative marking. Students are required to answer using OMR (Optimal Mark Recognition) sheets.

SECTION A

Unit 1: Time series analysis: Introduction, Uses and Importance, Components: Secular trend,

short-term variations, Random and irregular trends

Unit 2: Measurements of Trend: Graphic, Semi-average, least square and Moving Average, Merits and Demerits

Unit 3: Basics of Probability: Addition Law, Conditional probability, Multiplication law

SECTION B

Unit 4: Probability Distribution: Binomial distribution and Poisson distribution

Unit 5: Normal distribution- Meaning, Properties and fitting

Unit 6: Interpolation and Extrapolation.

Note: Statistical analysis should also be taught with the help of MS Excel, SPSS, or any other related software tool.

Suggested Readings

- A.M Goon, M.K Gupta and B. Dasgupta, fundamental of statistics Vol-I, World press Calcutta
- Anderson, D.R.; Sweeney, D.J. and Williams, T.A., "Statistics for Business and
- Economics", 2nd edition (2011), Thompson, New Delhi.
- Gupta SC: Fundamental of statistics, S. Chand & Company. New Delhi
- Gupta, SP: Statistical Methods, S. Chand & Company. New Delhi
- Kothari, C. R., "Research Methodology", 2nd Edition (2008), New Age International.
- Meyer, P.L. (1990): Introductory Probability and Statistical Applications, Oxford & IBH Pub.
- Monga, GS: Mathematics and Statistics for Economics, Vikas Publishing house, New Delhi.
- Rohatgi, V. K. and Saleh, A.K.M.E. (2010): An Introduction to Probability Theory and Mathematical Statistics, Wiley Eastern.

CERTIFICATE/ DIPLOMA IN STATISTICAL ANALYSIS AND RESEARCH METHODOLOGY
SEMESTER II
SARM 4: TIME SERIES ANALYSIS AND PROBABILITY DISTRIBUTIONS

EDITOR AND COURSE CO-ORDINATOR- DR. PINKY SRA

SECTION A

UNIT NO.	UNIT NAME
Unit 1	Time series analysis: Introduction, Uses and Importance, Components: Secular trend, short-term variations, Random and irregular trends
Unit 2	Measurements of Trend: Graphic, Semi-average, least square and Moving Average, Merits and Demerits
Unit 3	Basics of Probability: Addition Law, Conditional probability, Multiplication law

SECTION B

UNIT NO.	UNIT NAME
Unit 4	Probability Distribution: Binomial distribution and Poisson distribution
Unit 5	Normal distribution- Meaning, Properties and fitting
Unit 6	Interpolation and Extrapolation

CERTIFICATE/ DIPLOMA IN STATISTICAL ANALYSIS AND RESEARCH METHODOLOGY

SEMESTER II

TIME SERIES ANALYSIS AND PROBABILITY DISTRIBUTIONS

UNIT 1: TIME SERIES ANALYSIS

STRUCTURE

1.0 OBJECTIVES

After studying the Unit, the learner will be able to:

- Define the Meaning of Time series Analysis
- Distinguish different types of fluctuations in the time series analysis
- Understand how time series analysis is useful for forecasting
- decompose a time series into its various components
- Components of time series

1.1 INTRODUCTION

One of the important functions of business managers is to make forecasts about the future. This forecasting helps them in making the business decisions. Many Statistical Techniques help a business manager in forecasting the future. Time series analysis is one such technique. This technique is not only used by Business managers, rather other persons interested in forecasting also use this technique like economists, etc. Time series analysis is a tool with the help of which we try to predict future values based on data available to us. For example, if we have data on a company's sales for the last 10-12 years and we want to predict the likely sales of the company for the next year, we can do so using time series analysis.

Usually, the quantitative data of the variable under study are denoted by y_1, y_2, ...yn and the corresponding time units are denoted by t_1, t_2 .....tn. The variable 'y' shall have variations, as the values will show ups and downs. These changes account for the behaviour of that variable. Instantly it comes to our mind that 'time' is responsible for these changes, but this is not true. Because the time (t) is not the cause and the changes in the variable (y) are not the effect. The only fact, therefore, that we must understand is that several causes affect the variable and have operated on it during a given period. Hence, time becomes only the basis for data analysis.

Making decisions is aided by the ability to forecast any event. If we can comprehend the historical behaviour of that specific activity, forecasting becomes feasible. To comprehend historical behaviour, a researcher requires both historical data and a thorough examination of it. As a result, the necessity of time series analysis, time series fluctuations that account for changes in the series over time, and trend assessment for forecasting will all be covered in this unit.

Following are a few examples of time series analysis:

- A series of data related to production of goods, prices of goods or consumption level of goods.

- Data related to the rainfall or temperature of a region.
- The data related to sales profit etc of any business firm.
- The data related to exports and imports of the country.
- The data related to population, birth rate, or death rate in a country.

1.2 DEFINITION OF TIME SERIES

In time series we collect the data related to statistical observations and place such data in chronological order, that means in the order of occurrence of these observations. Based on these observations we can try to predict the future values of the observation. Following is the definition of time series analysis:

According to Ya-Lun-Chou "A time series may be defined as a collection of readings belonging to different periods of some economic variable or composite of variables".

According to W.Z. Hirsch, "The main objective in analyzing time series is to understand, interpret and evaluate change in economic phenomena in the hope of more correctly anticipating the course of future events".

1.3 UTILITY OF TIME SERIES ANALYSIS

The analysis of time series is of great utility not only to research workers but also to economists, businessmen and scientists etc., for the following reasons:

1. It helps in understanding the past behavior of the variables under study.
2. It facilitates forecasting future behavior with the help of the changes that have taken place in the past.
3. It helps in planning the future course of action.
4. It helps in knowing current accomplishments.
5. It is helpful to make comparisons between different time series and significant conclusions drawn therefrom.

Thus, we can say that the need for time series analysis arises in research because:

- to understand the behavior of the variables under study,
- to know the expected quantitative changes in the variable under study, and
- to estimate the effect of various causes in quantitative terms.

1.4 ESSENTIAL CONDITIONS OF TIME SERIES ANALYSIS

1. Time series analysis must consist of those values that are homogenous for example the sales data of every year must be in the same quantities as in kilograms. If sales of some years are given in quantity and others are given in value, then we cannot apply time series analysis.
2. The data present must be about time only. So, out of the two variables given, one variable should be time. For example, if a relation between Price and Demand is given it is not a time series.
3. The data must be arranged in chronological order.
4. The data must be available for a long time at least 10 to 12 years.
5. We must try to keep an equal gap between the two periods.
6. If the gap between the periods is not equal and some values are missing, we should try to find out those values using the interpolation.
7. The data must have some relation with the time. For example, if we are measuring the average marks of the students nn a class, it is not related to time.

1.5 ADVANTAGES OF TIME SERIES ANALYSIS

Time series analysis has several benefits for companies and scholars alike. Among its advantages are:

- **Data cleaning**: By removing noise and outliers, time series analysis techniques including smoothing and seasonality adjustments serve to improve the data's dependability and interpretability.
- **Understanding Data**: The fundamental structure of the data can be understood by models such as exponential smoothing or ARIMA. Understanding the true nature of the data can be aided by autocorrelations and stationarity measurements.
- **Forecasting**: Predicting future values from previous data is one of the main applications of time series analysis. For applications such as stock market analysis and company planning, forecasting is quite useful.
- **Finding Patterns and Seasonality:** Time series analysis can reveal in data underlying patterns, trends, and seasonality that may not be visible with bare eyes.
- **Visualisations:** Meaningful visualizations that demonstrate patterns, cycles, and abnormalities in the data can be produced using time series decomposition and other methods.

- **Efficiency**: Less data can occasionally be more when using time series analysis. Without being sucked into too complicated models or datasets, important insights can frequently be obtained by concentrating on key metrics and timeframes.
- **Risk assessment**: Decision-making processes related to finances and operations can be aided by modelling volatility and other risk factors over time.

1.6 COMPONENTS OF TIME SERIES

A large number of forces are there that affect the data. For example, if the sales of a company are changing over time, there are many forces responsible for it. We can classify these forces basically into four categories known as components or elements of the time series. The following are these components:

1. Secular Trend
2. Seasonal variations
3. Cyclical variations
4. Irregular variations

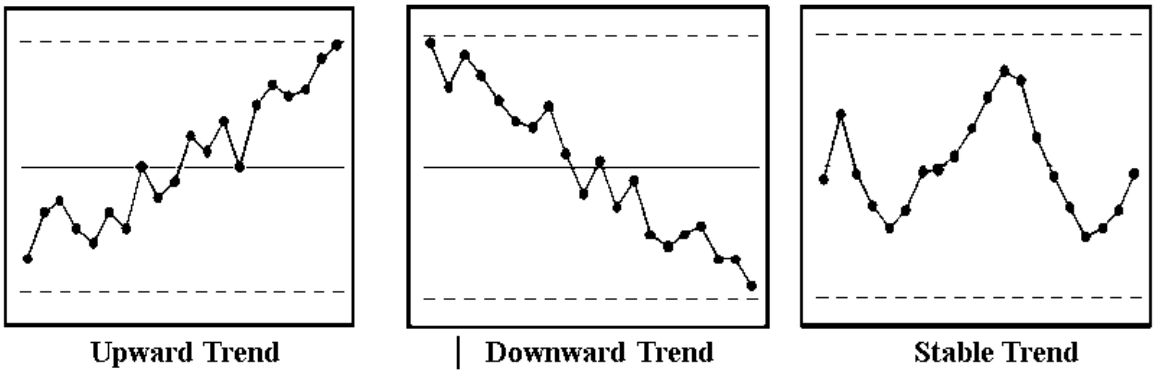
1.6.1 SECULAR TREND

The word *secular* is taken from the Latin word 'Saeculum', which means a ‘Generation’. So, as the name suggests, secular Trends are long-term Trends that normally occur over it period of 15 to 20 years. Sometimes, these trends may show upward results, and other times they may show downward results. For example, we can see that the number of persons who are traveling by air is increasing over a period of time. Similarly, we can see that the infant mortality rate in the country is decreasing over a period of time. These both are secular trends but one trend is showing an upward result and the other trend is showing a downward result.

Features of Secular Trend

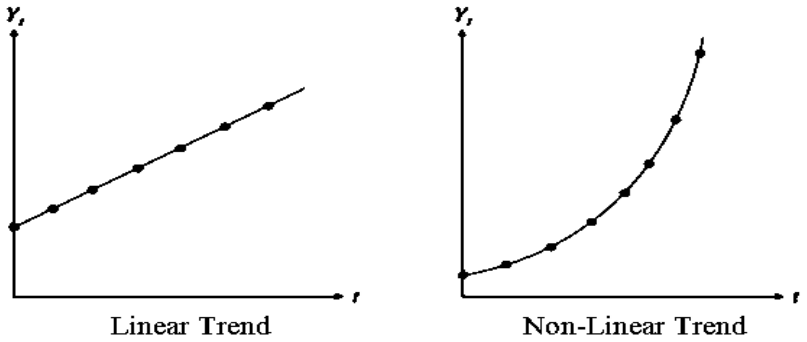
The following are the characteristics of secular trends:

1. These Trends are related to long periods.
2. These trends result from factors that are more or less stable. For example, the taste of people's change of Technology takes time and does not happen overnight.
3. These Trends may show upward, downward, or stable results.



Upward Trend | Downward Trend | Stable Trend

4. These trends may be linear or nonlinear. Linear Trends are those Trends that change proportionately over time and these are presented in a graph as a straight line. Nonlinear friends are those which does not change proportionately, so when we draw these trends on graph paper, these are not in a straight line.



Linear Trend | Non-Linear Trend

Uses of Secular Trend

The following are the benefits of studying secular trends:

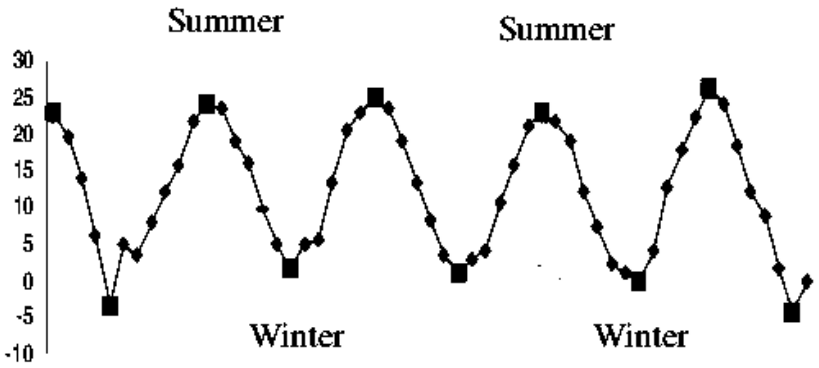
1. As these trends are long-term, they help us to understand the direction of change. We can find whether the phenomenon is increasing, decreasing, or is stable over time.
2. These Trends can help us in predicting the future.
3. Secular Trends can help us in comparing two or more series and we can see which series is changing more rapidly.
4. It can be used to extrapolate future values.
5. With the help of this trend, we can also study the impact of other components of time series.

1.6.2 SEASONAL VARIATIONS

Seasonal variations are short-term variations. These variations occur regularly and their trend is repetitive. These variations may occur every year, half-yearly basis, monthly basis, weekly basis, or any other period basis. There may be many reasons for these variations but these variations

generally occur due to the following two reasons:

1. **Climatic conditions**: Sometimes seasonal variations take place due to climate change. We can see that there are climatic cycles that occur during the year. This climatic cycle also effects on many things like the sales of a company, consumption patterns, etc. For example, in the rainy season sales of umbrellas increase, in the summer season sales of air conditioners increase and similarly during the winter season sales of Woolen clothes increase. These variations take place every year.
2. **Customs and traditions**: Sometimes seasonal variations take place due to customs and traditions. For example, in India is a tradition of purchasing new items in the household at the time of the Diwali festival. So, this is also a seasonal variation that takes place every year.



Features of Seasonal Variations

The following are characteristics of seasonal variations:

1. These variations are short-duration variations.
2. These variations repeat periodically.
3. It may have both an upward and downward trend, for example in winter sales of woollen garments increase but at the same time sales of soft drinks decrease.
4. These variations may occur on a yearly, quarterly monthly or weekly basis.
5. As these variations are repetitive and short-duration in nature, these are comparatively easy to analyze.

Uses of Seasonal Variations

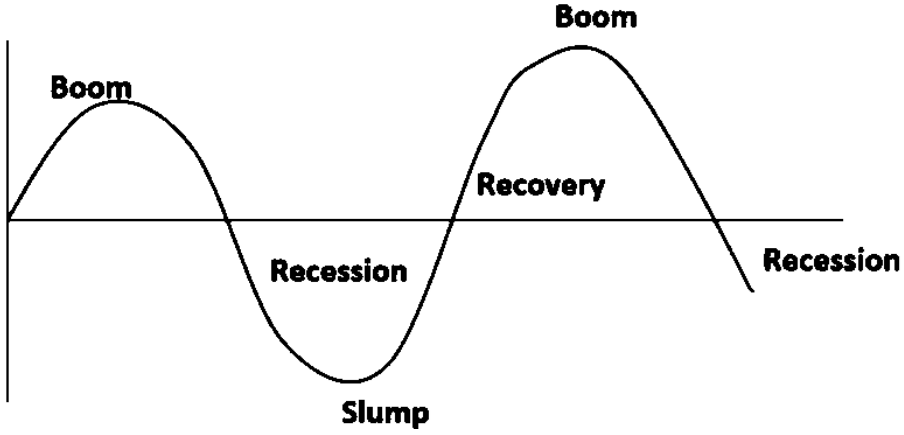
Following are the uses of seasonal variations

1. Analysis of these variations is very important for businesses in planning their production schedule. Business can decide their production according to seasonal variations.

2. Seasonal variations are useful for consumers also as they can plan their purchases according to the season. They know in advance which season is coming, so they can purchase items related to that season.
3. Seasonal variations also help consumers in making a bargain. Consumers may get off-season items at lower prices.
4. These variations help us to separate the effect of cyclical and irregular variations.
5. Businesses can use seasonal variations in many decisions like purchasing, inventory control, recruitment of employees, advertising, etc.
6. These variations help us in making short-term forecasts.

1.6.3 CYCLICAL VARIATIONS

As the term 'cycle' suggests, these variations are recurrent. These variations are long Run variations and show a recurring pattern of rise and decline. These variations are also known as oscillating movements. These variations do not have any fixed duration. Sometimes one cycle may be complete in 2-3 years, but some other times it may take 7-8 years to complete. For example, a business cycle is a cyclical variation that has four phases Boom, recession, depression, and recovery.



Features of Cyclical Variations

The following are features of cyclical variations:

1. Normally these occur over a long period that is more than 1 year.
2. There is no fixed duration of these variations, time one cycle is completed in 3 years but some other time it may take 7 years.
3. These variations are oscillating in pattern.

4. These are comparatively more difficult to measure.

Uses of Cyclical Variations

Following are uses of cyclical variations

1. Cyclical variations can help a business in Planning its strategy. Businesses can plan strategies according to the Boom or depression in the market.
2. Analysis of these variations can help businesses in predicting the turning points of cyclic variations.
3. Analysis of these variations can help business planning stabilization policies like diversification etc.
4. Analysis of these variations can help us in finding irregular variations.

1.6.4 IRREGULAR VARIATIONS

From the names of these variations, it is clear that these variations do not have any definite pattern and are irregular. These variations do not have any fixed time and occur due to accidental or random factors like strikes, floods, pandemics, wars, earthquakes, etc.

Features of Irregular Variations

The following are features of irregular variations.

1. These variations do not have any fixed pattern.
2. Mostly these variations are short duration.
3. These variations are very difficult to predict.
4. These occur due to random or accidental such as floods quakes wars etc.

1.7 DE-COMPOSITION OF TREND

As we have discussed above, any time series data comprises of various components namely Secular trends, seasonal variations, cyclical variations, or irregular variations. In the time series Analysis, we try to identify various components of time series separately. This can be done by measuring the impact of one component while we keep the other component constant. This process of finding each of the elements of time series separately is known as the De-composition of time series. There are many models which are normally used to analyze the time series. These are:

1.7.1 Additive Model

This model of decomposition assumes that the four elements of the time series are not dependent on each other and do not affect each other. Each trend operates independently. So, if we have to

measure the overall trend of the time series, it is a combination of all four elements. By adding the effect of all the elements, we can get the overall time series trend. Mathematically we can say that

$$\mathbf{Y=T+S+C+I}$$

$$\mathbf{Short\ term\ fluctuations = Y - T = S+C+I}$$

$$\mathbf{Cyclical\ and\ Irregular\ Fluctuation = Y - T - S = C + I}$$

$$\mathbf{Irregular\ Fluctuation = Y - T - S - C = I}$$

Where,

Y = time series value,

T = Secular Trend Variations,

S = Seasonal Variations,

C = Cyclical Variations and

I = Irregular Variations.

In the additive model, we assume that all the elements operate independently, but in reality, it is not true as all the elements have significant effects on each other and this is the major limitation of the additive model.

1.7.2 Multiplicative Model

The Multiplicative model is based on the assumption that all the components of the time series are related to each other and have significant effects on each other. So, if we want to calculate overall trend, it cannot be calculated by simply adding the four components. Rather it is multiple effect of all the four elements. So according to this model overall trend is

$$\mathbf{Y = T \times S \times C \times I}$$

$$\mathbf{Short\ term\ fluctuations} = \frac{Y}{T} = \mathbf{S \times C \times I}$$

$$\mathbf{Cyclical\ and\ Irregular\ Fluctuation} = \frac{Y}{T \times S} = \mathbf{C \times I}$$

$$\mathbf{Irregular\ Fluctuation} = = \frac{Y}{T \times S \times I} = \mathbf{I}$$

Here it is important to mention that the values of S, C, and I are not absolute values rather these are relative variations and these are expressed in relative change or some indices.

The multiplicative model is typically more appropriate and regularly used in business research to analyze time series. Since various factors interact to produce data linked to business and economic time series, no single element can be held accountable for producing a particular sort of variance.

1.7.3 Difference between Additive Model and Multiplicative Model

Additive Model	Multiplicative Model
1. It is based on the assumption that all elements of a time series are independent of each other.	It is based on the assumption that all elements of a time series are dependent on each other.
2. Under this model the overall trend can be found by adding the four elements.	Under this model, the overall trend can be found by multiplying the four elements.
3. Under this model absolute values of the four elements are taken for calculating the overall trend.	Under this model, relative values of the four elements are taken for calculating overall trend.

1.8 PRELIMINARY ADJUSTMENTS

Before we proceed with the task of analysing time series data, it is necessary to make relevant adjustments to the raw data. They are:

1. **Calendar variations:** As we are aware, all the calendar months do not have the same number of days. For instance, the production in February may be less than other months because of fewer days and if we take the holidays into account the variation is greater. Therefore, adjustments for calendar variations have to be made.
2. **Price changes:** As price level changes are inevitable, it is necessary to convert monetary values into real values after taking into consideration the price indices. This is the process of deflating.
3. **Population changes:** Population grows constantly. This also calls for adjustment in the data for the population changes. In such cases, if necessary, per capita values may be computed (dividing original figures by the total population).

1.9 TIME SERIES ANALYSIS'S CHALLENGES

Although time series analysis offers many benefits, it also has drawbacks and difficulties of its

own, including:

- **Restricted Scope**: Only time-dependent data can be used for time series analysis. Cross-sectional or strictly categorical data are not appropriate for it.
- **Noise Introduction**: Data may become noisier as a result of techniques like differencing, which could mask underlying patterns or trends.
- **Interpretation Challenges**: To make it simpler to comprehend the practical ramifications of the findings, some modified or differed values may require more intuitive interpretation.
- **Problems with Generalisation**: When an analysis is based on a particular, isolated dataset or period, results may not always be generalizable.
- **Model Complexity**: Choosing the right model can have a significant impact on the outcomes, and using the wrong model might provide inaccurate or misleading findings.
- **Non-Independence of Data**: Time series data points are not necessarily independent, which might create bias or mistake in the study, unlike other types of statistical analysis.
- **Data Availability**: To get accurate results, time series analysis frequently needs a large number of data points, which are not always readily available or accessible.

1.10 SUM UP

- Time series analysis is a situation where there are two variables in the problem and out of that one variable is necessarily the time factor.
- This analysis is a very useful tool for forecasting.
- over time there are fluctuations in the items.
- These fluctuations are mainly due to four factors called components of time series.
- These components are Secular trends, seasonal variations, cyclical variations and irregular variations.
- There are two models of time series, these are additive models and multiplicative models.

1.11 Key Terms

- **Time Series:** In time series we collect the data related to statistical observations and place such data in chronological order, that means in the order of occurrence of these observations. Based on these observations we can try to predict the future values of the observation.

- **Secular Trend:** Secular Trends are long-term Trends that normally occur throughout 15 to 20 years. Sometimes, these trends may show upward results and other times they may show downward results.
- **Seasonal Variations:** Seasonal variations are short-term variations. These variations occur regularly and their trend is repetitive. These variations may occur yearly, half-yearly basis, monthly basis, weekly basis or any other time basis. These may occur due to climatic conditions or due to customs and traditions.
- **Cyclical Variations:** These variations are long-run variations and show a recurring pattern of rise and decline. These variations are also known as oscillating movements. These variations do not have any fixed duration. Sometimes one cycle may be complete in 2-3 years, but other times it may take 7-8 years to complete for example Trade cycles.
- **Irregular Variations:** From the name of these variations, it is clear that these variations do not have any definite pattern and are irregular. These variations do not have any fixed time and occur due to accidental or random factors like strikes, floods, pandemics, war, earthquakes, etc.
- **Additive Model of Time Series:** This model of decomposition assumes that the four elements of time series are not dependent on each other and do not affect each other. Each trend operates independently. So, if we have to measure the overall trend of the time series, it is a combination of all four elements.
- **Multiplicative Model:** The Multiplicative model is based on the assumption that all the components of the time series are related to each other and have significant effects on each other. So, if we want to calculate the overall trend, it is multiple effect of all the four elements.

1.12 QUESTIONS FOR PRACTICE

Q1. What is time series? Give its significance and limitations.

Q2. What is the Utility of time series analysis

Q3. Explain the Essential Conditions of Time Series Analysis

Q4. Give the Advantages of Time Series Analysis

Q5. Explain the Decomposition of Trend

Q6. What are the components of time series?

Q7. Give different types of trends in time series.

Q8.Give multiplicative and additive models of time series.

1.13 FURTHER READINGS

- J. K. Sharma, *Business Statistics,* Pearson Education.
- S.C. Gupta, *Fundamentals of Statistics,* Himalaya Publishing House.
- S.P. Gupta and Archana Gupta, *Elementary Statistics,* Sultan Chand and Sons, New Delhi.
- Richard Levin and David S. Rubin, *Statistics for Management,* Prentice Hall of India, New Delhi.
- M.R. Spiegel, *Theory and Problems of Statistics,* Schaum's Outlines Series, McGraw Hill Publishing Co.

CERTIFICATE/ DIPLOMA IN STATISTICAL ANALYSIS AND RESEARCH METHODOLOGY

SEMESTER II

TIME SERIES ANALYSIS AND PROBABILITY DISTRIBUTIONS

UNIT 2: MEASUREMENTS OF TREND

STRUCTURE

2.0 OBJECTIVES

After studying the Unit, learners will be able to know about:

- how time series analysis is useful for forecasting
- Free Hand Graphic Method
- Semi-Average Method
- Moving Average Method
- Least Square Method

2.1 MEANING OF TRENDS

Trend means the direction or tendency of a series of data over a long period. This tendency may be linear or non-linear and upward and downward type. There is a need to measure this component so that one can remove the effect of factors from the time series that are responsible for this. For example, an increase/decrease in the price of a commodity, shares, gold due to an economic boom or decline helps in understanding the relation between factors and also in future prediction. So that one can forecast the values of the study variable more exactly through modeling as well as a study of its characteristics.

2.2 MEASUREMENT OF TREND

We want to know whether the values are increasing over a period of time, decreasing over a period of time or these are stable over a period of time, known as Trend. We generally assume that the past behaviour of the data will continue in the future as well, so finding the trend could help us in predicting the future. There are many methods available for measurements of secular trends. Some of them are method of curve fitting by least square method and moving average method. These methods will be discussed one by one in this module along with their merits and demerits. Basically, there are four methods of finding the trend which are as follows:

- Free-hand graphic method
- Semi-average method
- Moving average method
- Method of least square

2.3 FREE HAND GRAPHIC METHOD

This is the simplest method of finding the trend and is very flexible. This method is also known as the ‘free hand curve fitting method’. Despite being extremely simple, a free-hand method is not

widely accepted because it produces various trend values for the same data depending on who is doing the work or even when the same person is doing it. It should be highlighted that because the free-hand method is so subjective, different researchers can draw different trend lines from the same set of data. As a result, using it as the foundation for forecasting is not advised, especially in cases when the time series shows extremely irregular movements.

The following are the steps for finding trend under this method:

1. In the graph paper line chart is to be drawn.
2. For this purpose, time is taken on the x-axis whereas values are taken on the y-axis.
3. Plot all the given values in the graph paper.
4. Then we join all the points in the graph paper to show the actual value.
5. After that smooth straight line is drawn which passes through the middle of the actual values drawn.
6. This line is the trend line.

2.3.1 Points to be considered in drawing Free Hand Graphic Method

Following are precautions that be taken while drawing a trend line

1. It should be a smooth line.
2. The number of points above the line and the number of points below the line should be equal.
3. If there are cycles in the data, the number of cycles above the line and number of cycles below the line should be equal.
4. We must try that the trend line should pass through the middle of the points.
5. We should try to keep sum of vertical distance between trend line and the points nearly zero, which means we must try to have minimum deviation.

2.3.2 Merits of Free Hand Graphic Method

Following are the benefits of the graphic method

1. It is simple to draw
2. It does not need any calculations
3. This is a very flexible method and is not affected by the fact that data is linear on or non-linear.
4. An experienced statistician can use this tool very effectively.

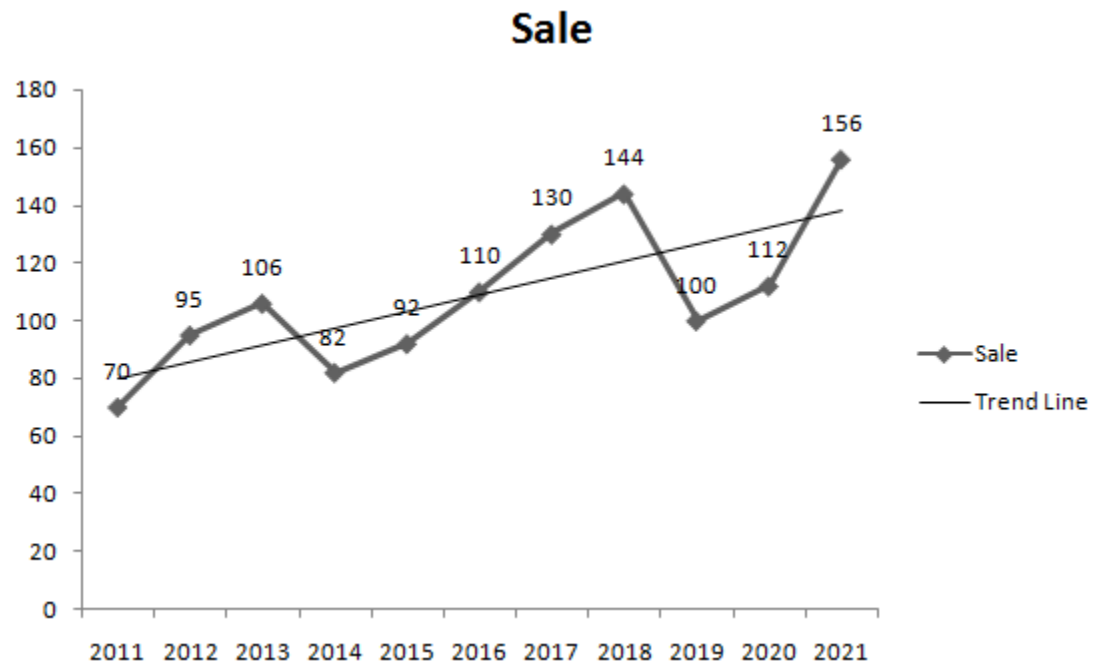
2.3.3 Demerits of Free Hand Graphic Method

1. This is a very crude method.
2. This method is very subjective and there are chances of personal biasness.
3. It needs a lot of experience to draw this chart.
4. With the change in scale of the graph there is a change in trend also.

Example 1: Fit the straight-line graphic curve from the following data:

Year	**2011**	**2012**	**2013**	**2014**	**2015**	**2016**	**2017**	**2018**	**2019**	**2020**	**2021**
Sale	**70**	**95**	**106**	**82**	**92**	**110**	**130**	**144**	**100**	**112**	**156**

Solution:



From the above graph we can predict any value with the help of trend line.

CHECK YOUR PROGRESS (A)

1. Following is the data of Harshit Ltd. draw a straight trend line using free hand graphic method.

Year:	2009	2010	2011	2012	2013	2014	2015	2016	2017
Sales (in '000 kg):	20	22	24	21	23	25	23	26	24

2. On basis of following data fit straight trend line using free hand graphic method.

Year	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021
Production	64	82	97	71	78	112	115	131	88	100	146

2.4 SEMI AVERAGE METHOD

Semi average method is the second method of finding the trend line. This is an objective method and is not merely based on guesswork. Under this method, it is very easy to find a trend line. Following are the steps of semi average method:

1. Divide the series in two equal parts, for example, if there are 10 values take 5 values in each part.
2. In case of number of values is an odd number, the middle value may be left and the remaining values can be divided into two parts. For example, if there are 11 values, the 6th value may be left and will have two parts having five values each.
3. Find the Arithmetic mean of both the parts.
4. These arithmetic means are called semi averages.
5. Now these semi averages are plotted in the graph as points against middle of each time period for which these have been calculated.
6. Join the points to find out straight line Trend.

2.4.1 Merits of semi average method

1. This is simple and easy to draw.
2. This is objective method and does not suffer from limitation of biasness.
3. As the line drawn is extendable on both sides, we can predict future values also.

2.4.2 Demerits of semi average method

1. This method is useful only for Linear trends.
2. This method is based on Arithmetic mean which is not a perfect average.

Even Number of Years

Example 3: From the data given below find semi average trend line and also find out trend values.

Year	**2011**	**2012**	**2013**	**2014**	**2015**	**2016**	**2017**	**2018**	**2019**	**2020**
Sale '000'	**80**	**61**	**76**	**73**	**62**	**50**	**45**	**65**	**55**	**35**

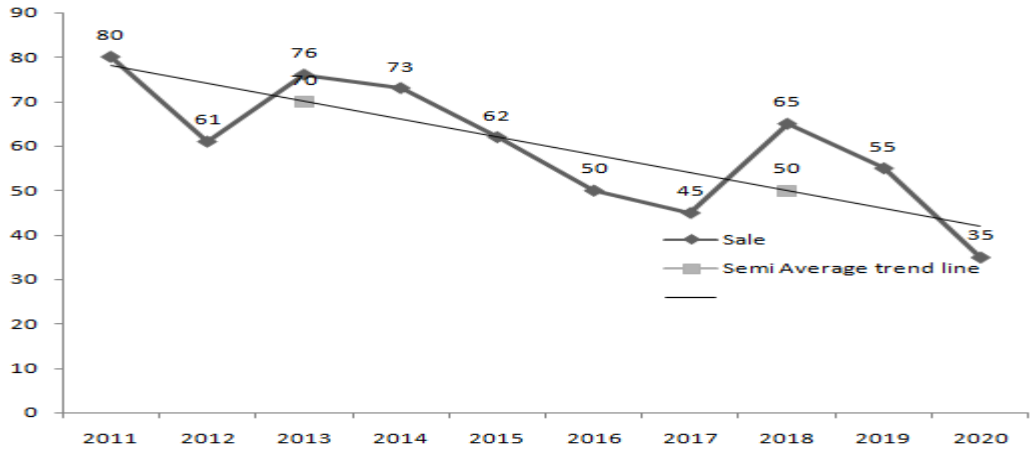
Solution:

As the number of years is even, we have got two blocks of five years each. Now we will find arithmetic mean of these two blocks and will write against middle of the block.

Year	Sale '000'	Semi Average

2011	80	
2012	61	
2013	76	$= \frac{80+61+76+73+62}{5} = \frac{350}{5} = 70$
2014	73	
2015	62	
2016	50	
2017	45	
2018	65	$= \frac{50+45+65+55+35}{5} = \frac{250}{5} = 50$
2019	55	
2020	35	

For finding the trend in the graph 70 is plotted against year 2013 and 50 is plotted against the year 2018.



$$\text{Annual increment} = \frac{Difference\ in\ Semi\ Average\ values}{Difference\ in\ two\ years\ to\ which\ Semi\ Average\ belongs}$$

$$\text{Annual increment} = \frac{50-70}{2018-2013} = \frac{-20}{5} = -4$$

As we can see from the above data that semi average is showing a downward trend so this annual increment will be deducted to semi average of 2013 onwards. For finding the values of the years before 2013 it will be added to the value every year. So, trend values are:

Year	Actual Sale '000'	Trend Sale '000'

2011	80	78 (74+4)
2012	61	74 (70+4)
2013	76	70
2014	73	66 (70-4)
2015	62	62 (66-4)
2016	50	58 (62-4)
2017	45	54 (58-4)
2018	65	50 (54-4)
2019	55	46 (50-4)
2020	35	42 (46-4)

Odd Number of Years

Example 4: From the data given below find semi average trend line and also find out trend values.

Year	**2011**	**2012**	**2013**	**2014**	**2015**	**2016**	**2017**	**2018**	**2019**	**2020**	**2021**
Sale '000'	**70**	**96**	**108**	**82**	**94**	**110**	**128**	**142**	**98**	**112**	**150**

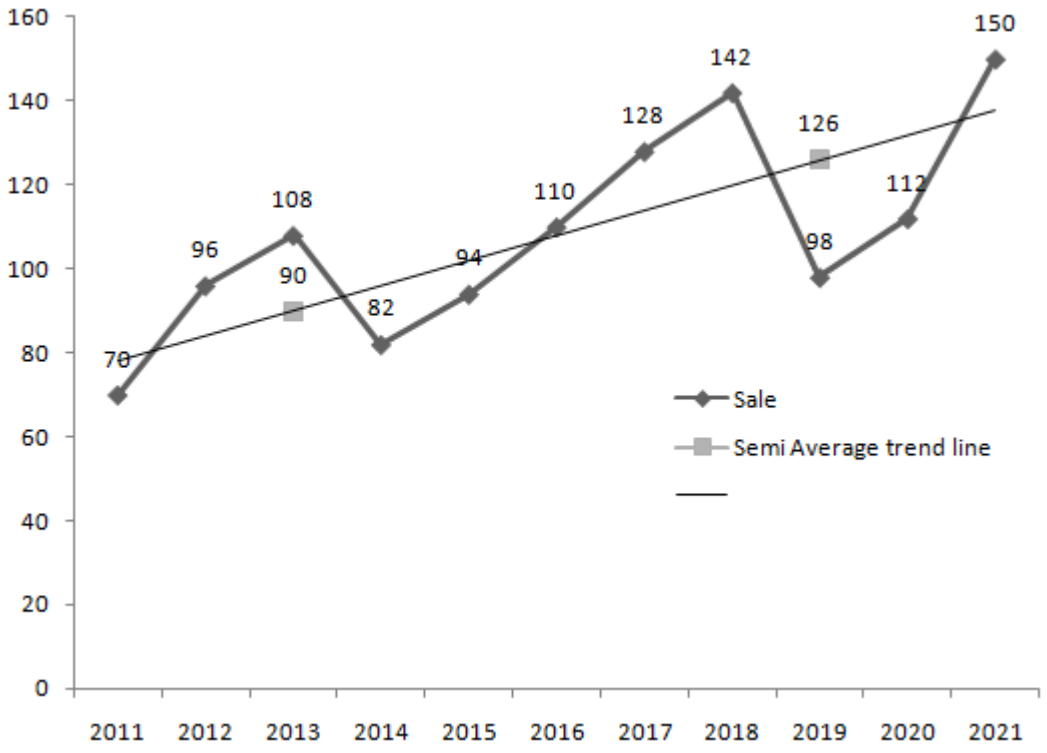
Solution:

As the number of years is odd, the middle year 2016 is left and we have got two blocks of five years each. Now we will find the arithmetic mean of these two blocks and will write against middle of the block.

Year	Sale '000'	Semi Average
2011	70	
2012	96	
2013	108	$= \frac{70+96+108+82+94}{5} = \frac{450}{5} = 90$
2014	82	
2015	94	

2016	110	
2017	128	
2018	142	
2019	98	$= \frac{128+142+98+112+150}{5} = \frac{630}{5} = 126$
2020	112	
2021	150	

For finding the trend in the graph 90 is plotted against year 2013 and 126 is plotted against the year 2019.



Annual increment = $\frac{Difference\ in\ Semi\ Average\ values}{Difference\ in\ two\ years\ to\ which\ Semi\ Average\ belongs}$

Annual increment = $\frac{126-90}{2019-2013} = \frac{36}{6} = 6$

As we can see from the above data that semi average is showing an upward trend so this annual increment will be added to semi average of 2013 onwards. For finding the values of the years before 2013 it will be deducted from the value every year. So, the trend values are:

Year	Actual Sale '000'	Trend Sale '000'

2011	70	78 (84-6)
2012	96	84 (90-6)
2013	108	90
2014	82	96 (90+6)
2015	94	102 (96+6)
2016	110	108 (102+6)
2017	128	114 (108+6)
2018	142	120 (114+6)
2019	98	126 (120+6)
2020	112	132 (126+6)
2021	150	138 (132+6)

CHECK YOUR PROGRESS (B)

1. The production of Mahanta Ltd fits a straight-line trend using a semi-average method:

Year	2011	2012	2013	2014	2015	2016	2017	2018
Production (‘000 Units)	200	210	218	192	204	216	224	228

Also, predict the value of 2020.

2. Fit straight line trend using Semi Average Method

Year	2012	2013	2014	2015	2016	2017	2018
Sales (in thousand units)	101	106	114	110	109	115	112

3. Sales of Abhinav are given, fit a straight-line trend using semi-average method:

Year	2012	2013	2014	2015	2016	2017	2018
Sales (‘000 Units)	80	90	92	83	94	99	92

Also, predict the value of 2021.

Answers:

1. 230, 3. 84

2.5 MOVING AVERAGE METHOD

Under this method, we try to find out the trend line using the concept of moving average. While considering matters such as trend of prices, sales, profits, etc., a particular type of average known as moving average is used. It is a measure of trend (long-term tendency of the data) in the

time series data. Moving average is an arithmetic average of data arising over a period of time and is calculated by replacing the first item in the average with the newly arising item. The successive averaging process does a smoothing operation in the time series data, i.c., it irons out fluctuations of uniform period and intensity. They can be eliminated by choosing the period of moving average that coincides with the period of the cycles i.e. periodic movements. Even if the periodic move with the period of the cycle i.e., periodic movements. Even if the periodic move merit is absent in the time series, the irregularities of data can be reduced to a large extent by the moving average process. If we choose this method, we should select a period for calculation. The period may be 3 years or 5 years or 6 years or 12 years etc., which is to be decided by considering the duration of the cycle.

For this, first of all, decide the period for which the moving average is to be calculated, for example, we can take the 3-year moving average, 4-year moving average, 5-year moving average or so on.

The following are the steps in this method:

1. First of all, decide the length of the period for which the moving average will be taken.
2. Calculate the moving average of the first group starting with first item.
3. After that find out the moving average of the second group leaving the first item.
4. Repeat this process until the moving average is calculated for all the groups ending with the last item.
5. Write the first moving average in front of the middle item of the group.
6. Repeat this process till all the moving averages are placed in front of the middle item of the group.
7. In case, an even number of years is taken as period of the moving average, the moving average is placed in middle of the period and then the average of the adjacent averages is placed against mid item.

2.5.1 Advantages of moving average

1. This method is easy to adopt.
2. It is a flexible method and any period can be taken as a moving average upon the period of cyclical Trend.
3. This method is free from bias.
4. Moving average reduces the impact of cyclical variations.

5. This method is not only useful for the measurement of trends but could also help in finding seasonal, cyclical and irregular variations.

2.5.2 Demerits of moving average method

1. This method cannot be used for predicting the future values.
2. We cannot calculate the trend for all the years as items beginning and some items at the end are lost.
3. It is very difficult to decide the period of moving average.
4. This Method is greatly affected by the presence of extreme values.
5. This method is not useful when we are estimating non-linear Trends.

Odd period Moving Average

Example 5: Calculate 3 yearly and 5 yearly moving averages for the following data:

Year	**2009**	**2010**	**2011**	**2012**	**2013**	**2014**	**2015**	**2016**	**2017**	**2018**	**2019**	**2020**
Sales	**52**	**49**	**55**	**49**	**52**	**57**	**54**	**58**	**59**	**60**	**52**	**48**

Solution:

Following are the steps for 3 yearly moving average

1. First, compute the total of value of first three years (2009, 2010, 2011) and place the three-year total against the middle year of 2010.
2. Now, leaving the first year's value, add up the values of the next three years (2010, 2011, 2012) and place the three-year total against the middle year 2011.
3. Repeat the process till last year's value i.e. 2020 is taken up.
4. Now divide the three-year's total by 3 to get the average and place it in the next column. All these values represent the required trend values for the given year.
5. The same process can be repeated for 5 yearly moving average.

Year	Sale	3-Year Moving Total	3-Year Moving Average	5-Year Moving Total	5-Year Moving Average
2009	52				
2010	49	156	52		
2011	55	153	51	257	51.4
2012	49	156	52	262	52.4
2013	52	158	52.7	267	53.4
2014	57	163	54.1	270	54

2015	54	169	56.3	280	56
2016	58	171	57	288	57.6
2017	59	177	59	283	56.6
2018	60	171	57	277	55.4
2019	52	160	53.3		
2020	48				

Even period Moving Average:

Example 6: Calculate 3 yearly and 5 yearly moving averages for the following data:

Year	**2011**	**2012**	**2013**	**2014**	**2015**	**2016**	**2017**	**2018**	**2019**	**2020**
Sales	**250**	**260**	**275**	**300**	**290**	**310**	**318**	**325**	**350**	**340**

Solution: Following are the steps for 4 yearly moving average

1. First, compute the total value of the first four years (2011, 2012, 2013, 2014) and place the four-year total in between the 2nd and 3rd years i.e. between 2012 and 2013.
2. Now, leaving the first year's value, add up the values of the next four years (2012, 2013, 2014, 2015) and place the total b 2011 between 2013 and 2014.
3. Repeat the process till last year's value i.e. 2020 is taken up.
4. Now divide the four years total by 4 to get the average and place it in the next column. All these values represent the required trend values for the given year.
5. Divide the first two four yearly averages by 2 to get the required trend values corresponding to the given years as shown in the table:

<table>
<tr><th>Year</th><th>Value</th><th>4 Yearly Total</th><th>4 Yearly Average</th><th>Trend Value</th></tr>
<tr><td>2011</td><td>250</td><td rowspan="2"></td><td rowspan="2"></td><td></td></tr>
<tr><td rowspan="2">2012</td><td rowspan="2">260</td><td rowspan="2"></td></tr>
<tr><td rowspan="2">1085</td><td rowspan="2">271.25</td></tr>
<tr><td rowspan="2">2013</td><td rowspan="2">275</td><td rowspan="3">276.25</td></tr>
<tr><td rowspan="2">1125</td><td rowspan="2">281.25</td></tr>
<tr><td rowspan="2">2014</td><td rowspan="2">300</td><td rowspan="3">287.5</td></tr>
<tr><td rowspan="2">1175</td><td rowspan="2">293.75</td></tr>
<tr><td rowspan="2">2015</td><td rowspan="2">290</td><td rowspan="3">299.12</td></tr>
<tr><td>1218</td><td>304.5</td></tr>
</table>

2016	310			307.63
		1243	310.75	
2017	318			318.25
		1303	325.75	
2018	325			329.5
		1333	333.25	
2019	350			
2020	340			

CHECK YOUR PROGRESS (C)

1. Calculate 3 yearly moving averages for the following data:

Year	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
Sales	11200	12300	10600	13400	13800	14500	11600	14300	13600	15400

2 Calculate 5 yearly moving averages for the following data:

Year	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
No. of Employees	332	317	357	392	402	405	410	427	405	438

3 Calculate 4 yearly moving averages for the following data:

Year	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
Value	100	105	115	90	95	85	80	65	75	70	75	80

Answers:

1. 11366.7, 12100, 12600, 13900, 13300, 13466.7, 13166.7, 14433.3
2. 360, 374.6, 393.2, 407.2, 409.8, 417
3. 101.875, 88.75, 91.875, 84.375, 78.75, 74.375, 71.875, 73.125

2.6 LEAST SQUARE METHOD

This is the most scientific and popular method of finding the trend line. Under this method, the lines of best fit are drawn as the lines trend. These lines are known as the lines of the best fit

because, with help of these lines we can estimate the values of variables according to the different time period. According to the Least Square method, trend line should be plotted in such a way that sum of square of the difference between actual value and estimated value of the dependent variable should be least or minimum possible. Mathematically this line is represented by

$Yc = a + bX$

Where Yc – Computed Trend Value

X – Independent Variable represented by time

a & b – Constants

2.6.1 Direct Methods to Estimate Trend Line

Following are steps for finding trend line with the help of the Direct Method:

1. Take the problem with two variables with X variable as time and other variables for which trend is to be computed like sales, population, etc represented by Y.
2. Assume first year as base year and put the value '0' against it.
3. Now put value 1 against second year, 2 against third year and so on till all the years are covered.
4. Now find the values of $\sum X$, $\sum X^2$, $\sum XY$ from the given values.
5. Put these values in following equation:

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

   $\Sigma XY = a \Sigma X + b\Sigma X^2$
6. Solve these equations simultaneously and find the values of 'a' and 'b'.
7. Put value of 'a' and 'b' in trend equation $Yc = a + bX$.
8. Now this trend equation can be used for finding the trend values.

Example 7. The data of sales of Alpha Ltd is given for last 9 years. On the basis of the data find trend value of the year 2021 using the method of least square.

Year	**2012**	**2013**	**2014**	**2015**	**2016**	**2017**	**2018**	**2019**	**2020**
Sales '000'	**10**	**12**	**15**	**20**	**30**	**40**	**50**	**60**	**70**

Solution:

Year	X	Sales (Y)	X^2	XY
2012	0	10	0	0
2013	1	12	1	12

2014	2	15	4	30
2015	3	20	9	60
2016	4	30	16	120
2017	5	40	25	200
2018	6	50	36	300
2019	7	60	47	420
2020	8	70	64	560
	$\sum X = 36$	$\sum Y = 307$	$\sum X^2 = 204$	$\sum XY = 1702$

This is given by $Y = a + bX$

where *a* and *b* are the two constants which are found by solving simultaneously the two normal equations as follows:

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

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

Substituting the given values in the above equations we get,

$307 = 9a + 36b$ …….. (i)

$1702 = 36a + 204b$ …….. (ii)

Multiplying the eqn. (i) by 4 we get

$1228 = 36a + 144b$…….. (iii)

Subtracting the equation (iii) from equation (ii) we get,

$1702 = 36a + 204b$

$-1228 = -36a - 144b$

$474 = 60b$

or $b = 7.9$

Putting the above value of b in the eqn. (i) we get,

$307 = 9a + 36(7.9)$ or

$9a = 307 - 284.4$ or

$a = 2.51$

Thus, $a = 16$, and $b = 1.6$

Putting these values in the equation $Y = a + bX$ we get

$Y = 2.51 + 7.9X$

So, if we want to calculate the trend value of the year 2021 the value of X will be 9 (as 2012 is our base year and its value is 0), the value of Y will be

Y = 2.51+ 7.9 (9) = 73.61

2.6.2 Short cut Method

In the direct method we take starting year as the base year. But in case we take the middle period as base yare we can save lot of time and calculation because when middle period is take as the base period the value of ∑X will be 0, hence the two simultaneous equations will become very easy in that case.

Equation (i) $\Sigma Y = na + b\Sigma X$

If $\Sigma X = 0$ than $\Sigma Y = na$

$a = \frac{\Sigma Y}{n}$

Equation (ii) $\Sigma XY = a \Sigma X + b\Sigma X^2$

If $\Sigma X = 0$ than $\Sigma XY = b\Sigma X^2$

$b = \frac{\Sigma XY}{\sum X^2}$

Odd number of Years

Example 8. The data of sales of Mahesh and Co is given for last 7 years. On the basis of the data find trend line using the method of least square and find trend value of 2021.

Year	**2014**	**2015**	**2016**	**2017**	**2018**	**2019**	**2020**
Sales '000'	**672**	**824**	**967**	**1204**	**1464**	**1758**	**2057**

Solution: Since the number of years is odd, 2017 is taken as base year with value 0 and one year is taken as one unit.

Year	X	Sales (Y)	X^2	XY
2014	-3	672	9	-2016
2015	-2	824	4	-1648
2016	-1	967	1	-967
2017	0	1204	0	0
2018	1	1464	1	1464
2019	2	1758	4	3516
2020	3	2057	9	6171

	∑X = 0	∑Y =8946	∑X ^{2}= 28	∑X Y =6520

As ∑X is 0, we can apply short cut method

$$a = \frac{\Sigma Y}{n} = \frac{8946}{7} = 1278$$

$$b = \frac{\Sigma XY}{\sum X^2} = \frac{6520}{28} = 232.9$$

Putting these values in the equation Y = *a* + bX we get

Y = 1278+ 232.9 X

So, if we want to calculate the trend value of the year 2021 the value of X will be 4 (as 2017 is our base year and its value is 0), the value of Y will be

Y = 1278+ 232.9 (4) = 2209.6

Odd number of Years

Example 9. The data of sales of Abhilasha Ltd is given for last 8 years. On the basis of the data find trend line using the method of least square and find trend value of 2021.

Year	**2013**	**2014**	**2015**	**2016**	**2017**	**2018**	**2019**	**2020**
Sales '000'	**80**	**90**	**92**	**83**	**94**	**99**	**92**	**104**

Solution: Since the number of years is even, so will take the origin as mid-point of 2016 and 2017 and further for sake of simplicity we one year is taken as two units (6 Month as 1 unit).

Year	X	Sales (Y)	X^2	XY
2013	-7	80	49	-560
2014	-5	90	25	-450
2015	-3	92	9	-276
2016	-1	83	1	-83
2017	1	94	1	94
2018	3	99	9	297
2019	5	92	25	460
2020	7	104	49	728
	∑X = 0	∑Y =734	∑X ^{2}= 168	∑X Y =210

As ∑X is 0, we can apply short cut method

$$a = \frac{\Sigma Y}{n} = \frac{7346}{8} = 91.75$$

$$b = \frac{\Sigma XY}{\sum X^2} = \frac{210}{168} = 1.25$$

Putting these values in the equation Y = *a* + bX we get

Y = 91.75+ 1.25 X

So, if we want to calculate the trend value of the year 2021 the value of X will be 9 (as mid of 2016 and 2017 is our base year and 1 year is taken as 2 units), the value of Y will be

Y = 91.75+ 1.25 (9) = 103

2.6.3 Merits of Least Square Method

1. There is no subjectiveness in this method as it is based on mathematical calculations.
2. This method is known as method of best fit, reason being the sum of deviations between trend and actual values is zero and sum of square of deviations is least.
3. This method can predict future values, that thing is not possible in moving averate.
4. This method gives us annual growth or decline rate also. The value of 'b' in the equation is growth of decline rate. If 'b' is positive than it is growth and if it is negative than it is decline.
5. It is based on all the values of the data.

2.6.4 Demerits of Least Square Method

1. This method involves lot of mathematical calculation, so is difficult for a layman.
2. This method finds trend value only and seasonal, cyclical and irregular variations are completely ignored.
3. If a new value is added to the data, we have to make the complete calculations once again.

CHECK YOUR PROGRESS (D)

1. These are the number of salesmen working in Alpha Ltd:

Year	2011	2012	2013	2014	2015	2016
Salesmen	28	38	46	40	56	60

Fit straight line trend using method of least squares.

2. Fit a straight-line trend by Method of least square and estimate the exports of 2021 using the short cut method:

Year	2013	2014	2015	2016	2017	2018	2019	2020
Exports	15	20	24	29	35	45	60	85

3. Determine the equation of straight line which best fits the following data

Year	2012	2013	2014	2015	2016	2017	2018	2019	2020
Value	620	713	833	835	810	745	726	806	861

4. Determine the equation of straight line which best fits the following data

Year	2001	2002	2004	2006	2007
Sales 'Lacs'	5	8	12	20	25

5. Determine the equation using the method of least square from a number of accidents from the following data and find trend values also.

Year	2001	2002	2003	2004	2005	2006	2007	2008
Accidents	38	40	65	72	69	60	87	95

Answers

1. Y = 44.67 + 2.97 X,
2. Y = 39.125 + 4.517 X; value of 2021 – 79.778
3. Y = 709.51 + 15.65 X,
4. Y = 14 + 3.23 X (taking 2004 as year of origin.
5. Y = 65.75 + 3.667 X (taking 2004.5 as year of origin).

Trend Values 40.081, 47.415, 54.749, 62.083, 69.417, 76.751, 84.085, 91.419

2.7 SUM UP

In this unit, discussed how to measure secular trend in the data using different methods. A set of quantitative data arranged on the bast of *TIME* are referred to as *Time Series*. The analysis of time series is done to understand the dynamic conditions for achieving the short-term and long-term goals of institutions. With the help of the techniques of time series analysis the future pattern can be predicted on the basis of past trends. In literature, there are many methods available for measurements of secular trend. Some of them are method of curve fitting by least square method and moving average method. These methods are discussed in this module along with their merits and demerits.

2.8 QUESTIONS FOR PRACTICE

Q1.What is free hand curve method?

Q2.What is semi average method of time series?

Q3.How predictions are made using method of least square.

Q4.What is moving average trend. How it is determined.

Q5.Give various methods of estimating trend along with their respective merits and limitations.

2.9 SUGGESTED READINGS

- Gupta, S. C. and Kapoor, V. K., *Fundamentals of Applied Statistics*, Sultan Chand & Sons, New Delhi, 2009.
- Gupta, S.C., *Fundamentals of Statistics,* Himalaya Publishing House.
- Richard Levin and David S. Rubin, *Statistics for Management,* Prentice Hall of India, New Delhi.
- Gupta, S. P., *Statistical Methods*, Sultan Chand & Sons, New Delhi, 2012.
- Sharma, J. K., *Business Statistics,* Vikas Publishing House, 2014.

CERTIFICATE/ DIPLOMA IN STATISTICAL ANALYSIS AND RESEARCH METHODOLOGY

SEMESTER II

TIME SERIES ANALYSIS AND PROBABILITY DISTRIBUTIONS

UNIT 3: BASICS OF PROBABILITY: ADDITION LAW, CONDITIONAL PROBABILITY, MULTIPLICATION LAW

STRUCTURE

3.0 OBJECTIVE

After reading this unit, learners will know about:

- Basic concepts of probability
- Different approaches to Probability
- Law of addition and multiplication
- Conditional Probability

3.1 INTRODUCTION

Words like probability, probable, chance, likelihood, etc. are frequently employed in casual speech. You may understand the meaning of these phrases in general. For instance, we might hear something like "there's a chance we win the cricket match today," or "the train might arrive late." It denotes a lack of certainty on the event's likelihood of occurring. The world in which we live makes it impossible for us to predict the future with absolute confidence. We study and apply probability because we have to deal with uncertainty. The definition of probability in statistics establishes it and has nothing to do with beliefs.

Probability theory has become one of the fascinating subjects in recent years. With the pioneering work of Jacob Bernoulli (1654-1705), Thomas Bayes (1702-1761), Joseph Lagrange (1736-1813), De Moivre (1667-1754) the theory of probability came into existence in the seventeenth century. Galileo (1564-1692) an Italian mathematician while dealing with some problems related to the theory of dice in gambling attempted to predict a quantitative measure of probability.

Further, Pierre Simon and Laplace (1749-1827) proposed their ideas and developed the first general theory of probability. After extensive research over some years finally published 'Theoretic analytique des probabilities' in 1812. Many Russian mathematicians made great contributions to the modern theory of probability. Chebychev (1821-94) founded the Russian school of statistics. Khinchine introduced the concept of the law of large numbers. Liapoun proposed the famous central limit theorem. Further, A. Kolmogorov axiomised the calculus of probability. A book on 'Foundations of Probability' was published by A. Kolmogorov in 1933. Initially, the probability theory was implemented at the gambling tables. Further, it was used to tackle social, economic, political and business problems. Today the concept of probability has assumed great importance and the mathematical theory of probability has become the basis for statistical applications in both social and decision-making situations.

In our day-to-day life, we face uncertainty and use probability theory, for decision-making. Probability constitutes the foundation of statistical theory and application. Information regarding probabilistic methods has become increasingly essential in the quantitative analysis of business and economic problems. The probability measure provides the decision maker with the means of quantifying the uncertainties that affect his choice of appropriate actions. It is extensively used in the quantitative analysis of business and economic problems. It is an essential tool in statistical

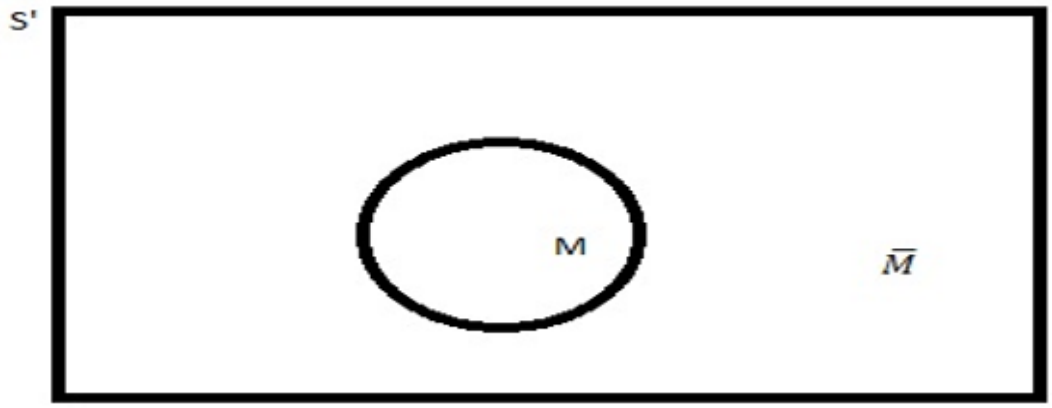
inference and forms the basis of decision-making.

Probability plays a significant part in statistics. It measures the likelihood of an event. One can observe different practical situations where prediction is quoted. For instance, the weather forecast quotes that there is a 90% chance of rain today. Probability helps us to make decisions in uncertain situations. It has important applications in all disciplines such as physics, chemistry, education, economics, etc. The probability theory has the purpose of providing mathematical models of situations affected or even governed by chance effects. One cannot predict with complete certainty the occurrence of the outcome of interest in any of the experiments. Broadly in the theory of probability, there are three possible states of expectation which are certainty, impossibility and uncertainty. The probability theory describes certainty as 1, impossibility as 0 and uncertainty lies between 0 and 1.

3.2 BASIC CONCEPTS OF PROBABILITY

1. **Universe**: The totality of a problem under consideration is known as a Universe, for example, the population of a country, the number of stars, All cards of a pack of cards, etc.
2. **Sample**: A part of totality is known as a Sample. For example, on students from a class ball from a basket, etc.
3. **Random experiment**: An experiment, in which the result is not predictable. In this type of experiment, the outcome depends upon chance.
4. **Trial**: Trial is defined as when a random experiment is carried.
5. **Event/Outcome**: An outcome of a random experiment is known as an event. For example, a coin tossed in the air is an experiment and a coin land with its head up is the outcome of the experiment. A parachute jumps from a plane is an experiment and it will fall as its outcome. The outcome of a random experiment which entails the occurrence of an event A is known as a favorable outcome to A.
6. **Exhaustive cases**: The total number of possible outcomes of a random experiment is called exhaustive cases. E.g. In tossing a dice total no. of outcomes is 6 i.e., 1,2,3,4,5,6.
7. **Favourable cases**: The total number of outcomes of a random experiment that confirms the happening of an event.
8. **Mutually Exclusive cases**: Any two or more events are called mutually exclusive if the occurrence of one of the events doesn't affect the occurrence of the other. E.g. In throwing of a dice, the set of all possible outcomes is mutually exclusive

9. **Equally probable cases**: The possible outcomes are said to be equally likely if none of the cases is preferred as compared to others. For instance, in tossing a coin the occurrence of head and tail are equally probable.
10. **Independent events:** Cases are said to be independent if the occurrence of one event is not affected by the occurrence of the other event. For eg, picking two balls from a bag containing 'a 'red balls and 'b' white balls is a trial and the occurrence of both red balls, both white balls and one red and second ball white are independent events.
11. **Simple event:** An event that includes one and only one of the outcomes for a random experiment is called a simple event.
12. **Compound event:** A compound event consists of more than one outcome.
13. **Complementary event:** Let 'M' represent the event of the number of favorable responses in the experiment and $\overline{M}$ represent the complementary event. It is defined as the number of non-favorable cases in the experiment. The representation of an event and its complement in the Venn diagram is discussed below:



For instance, a group of 2000 taxpayers, 400 have been audited by the IRS at least once. If one taxpayer is randomly selected from this group. The two complementary events for this experiment and their probability are

P(M) = 400/2000

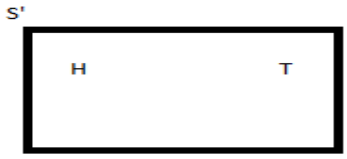
P ($\overline{M}$) = 1600/2000

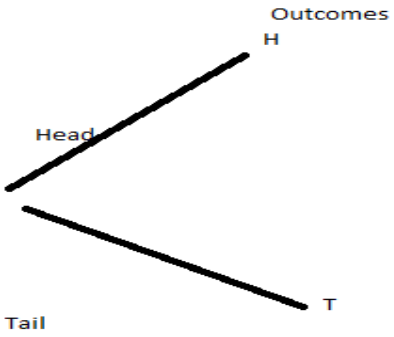
Where 'M' represents a taxpayer who has been audited by the IRS at least once $\overline{M}$ is the selected taxpayer never been audited

14. **i) Odd in favor:** Let $P = A/N$ is the probability that the event 'M' occurs and $Q = B/N$ be the probability that the event 'M' does not occur where $A + B = N$. If $p \geq 1/2$ then odd in favor of 'M' is the ratio A:B

    ii) Odd against favor: If $p \leq 1/2$, odd against A is defined as B: A

15. **Venn diagram-** It is a pictorial representation in the form of a rectangle, square or circle that predicts all the possible outcomes of an experiment
16. **Tree diagram-**It is the diagram in which each outcome is represented by a branch of tree. For example, Venn and Tree diagram of tossing a coin once if given as





where sample space = {H, T}, H represents head, T as tail.

17. **Intersection of events**- It gives us the common outcomes
18. **Joint Probability**- The probability of the intersection of two events is called joint probability P (M and N)
19. **Sampling with replacement**- This object was drawn at random is placed back into the given set and the set is mixed thoroughly. Then we draw the next object at a random.
20. **Sampling without replacement**-In this the object that was drawn is put aside

Definition of Probability

According to Beri (2007), "Probability is the chance that a particular event will occur." (What is the chance of getting a head when a coin is tossed.). To take another example, A company has launched a new product what is the chance that it will be successful?

According to Levin and Fox (2006): "The term probability refers to the relative likelihood of occurrence of any given outcome or event."

Probability of an outcome or event = The total number of times the occurrence of the event / the

total possible times an event can occur.

3.3 APPROACHES TO PROBABILITY

1. Classical approach
2. Empirical approach
3. Axiomatic approach
4. Personalistic approach

1. Classical approach/ a priori approach:

This approach involves the assumption that all the possible outcomes of an experiment are mutually exclusive and equally likely. If 'e' represents the event then the probability of occurrence of event 'e' is defined as

P(e)= [No of outcomes in favor of occurrence of event 'e'/Total no of outcomes]
= m/m+n

Since P(e) is a measure, its value lies between $0 \leq P(e) \leq 1$. If P (e^c) represents nonoccurrence of the event 'e' then P(e)+ If P (e^c)=1.

Example- Find the probability of obtaining a head and the probability of obtaining tail from a one toss of a coin.

Ans P(H)=1/2, P(T)=1/2

The probability calculations were based entirely upon logical prior reasoning.

Some drawbacks of the classical approach

1. It fails when the number of possible outcomes is infinitely large
2. It involves the feasibility of subdividing the possible outcomes of the experiment into mutually exclusive, exhaustive and equally likely cases

- **Law of large numbers**- If an experiment is repeated again and again the probability of an event obtained from the relative frequency approach the actual or theoretical probability
- **Counting rule**- an experiment consists of a total of 3 steps, the first step results in 'm' outcomes, the second step in 'n' outcomes and 3rd step in 'k' outcomes

Then total outcomes= m.n.k

For example**,** suppose we toss a coin three times. This experiment has three steps the first toss,

second toss and third toss. Each step has two outcomes a head and a tail. Therefore, the total outcomes=2.2.2=8. The outcomes are {HHH, HHT, TTH, TTT, THT, HTT, HTH, THH}

2. Relative frequency/Empirical approach:

It has been observed that the classical approach is quite effective for dealing with problems involving the game of chance. There are some major drawbacks that one analyzes while dealing with some specific types of problems. i)If one is asked to calculate the probability that a man aged 55 will die within 6 months.

ii) a production process used by a particular firm will produce a defective item.

In such situations, it is not feasible to justify mutually exclusive outcomes. To deal with such type of situations probability can be defined as

$$P(e) = \lim_{N\to\infty} M/N,$$

where an experiment is carried out N-times under the same homogeneous conditions and M are the outcomes. The relative frequency theoreticians agree that the only valid procedure for determining event probabilities is through repetitive experiments. R.A. Fisher, Von-Mises gave the concept of an empirical approach to the theory of probability through the notion of sample space.

3. Axiomatic approach:

It has been observed that in the classical approach, the estimation of probability depends upon mutually exclusive and equally likely events. In the case of the relative frequency approach/empirical approach, it involves a probabilistic problem that must be performed experimentally in the long duration of time under the same homogeneous conditions. The axiomatic approach does not restrict itself to the conditions imposed by both approaches. The axiomatic approach is a wise attempt to construct a theory of probability, largely free from the inadequacies of both the classical and empirical approaches. The main objective of this approach lies in the fact it makes available to the inquisitive mind a large number of abstract mathematical concepts, tools, and techniques. To define the probability in this approach we will first define the following terminology.

- **Sample Space:** It is defined as the set of all possible outcomes of a random experiment or the set of all exhaustive cases of the random experiment. If $a_1, a_2, \ldots a_n$ are all mutually exclusive possible outcomes of an experiment. The set $S' = \{a_1, a_2, \ldots a_n\}$, is known as the sample space of the experiment. Probability is defined as P(M)= n(M)/n(S')

where S' is the sample of a random experiment with a large number of trials, n(M) represents many sample points that are in favor of the event M and n(S') is the sample points in S'.

- **Marginal probability-** It is the probability of a single event without consideration of any other event. It is also known as simple probability

4. Personalistic approach-

A personalistic or subjective approach is defined as the degree of degree of confidence placed in the occurrence of an event by a particular individual based on the evidence available to him. This evidence may consist of relative frequency to data and any other quantitative or qualitative information.

Practical Problems of Probability

Ex 1. An unbiased coin is tossed. What is the probability that there will be a head?

Sol. When we toss a coin, then total (n) outcomes are = 2, (H, T)

Number of Favourable cases = m = 1, (H)

Required Probability. P(H)= $\frac{m}{n} = \frac{1}{2}$

Ex 2. A coin is tossed successively three times. What is its probability of getting?

(i) Exactly 3 heads (ii) at most 2 heads. (iii) least 2 heads

Sol: When we throw a coin successively three times, then the total outcomes are 2^3 =8

Here sample space is S = (HHH HHT HTH THH HTT THT TTH TTT)

(i) Exactly 3 heads = HHH = m = 1

Required probability = P(3H) = $\frac{m}{n} = \frac{1}{8}$

(ii) At most 2 heads: Means 2 heads, 1 head or 0 head.

The Favourable cases are = HHT HTH THỊ HTT THT TTH TTT

So, m = 7

Required Probability = $\frac{7}{8}$

(iii) At least 2 heads: - Means 2 heads or 3 heads.

The Favourable cases are = HHH, HHT, HTH, THH

So, m = 4

Required Probability = $\frac{m}{n} = \frac{4}{8} = \frac{1}{2}$

Eg 3: Two dice are thrown together. What is the probability that the sum of numbers on the two faces is divisible by 2 or 3 and 3 or 4?

Sol. When we throw two dice together then the total outcomes are as:

6^2=36, m=36

The Sample Space is

11	12	13	14	15	16
21	22	23	24	25	26
31	32	33	34	35	36
41	42	43	44	45	46
51	52	53	54	55	56
61	62	63	64	65	66

a) Divisible by 2 or 3: The favorable cases are:

(11, 13, 15, 21, 12, 22, 24, 26, 31, 33, 35, 36, 42, 44, 45, 46, 51, 53, 54, 55, 62, 63, 64,66)

So, m=24

Required Probability = $\frac{m}{n} = \frac{24}{36} = \frac{2}{3}$

(b) Divisible by 3 or 4: The favorable cases are:

(12, 13, 15, 21, 22, 22, 24, 26, 31, 33, 35, 36, 42, 44, 45, 51, 53, 54, 62, 63, 66)

So, m=20

Required Probability = = $\frac{m}{n} = \frac{20}{36} = \frac{5}{9}$

3.4 ADDITION THEOREM OF PROBABILITY

- When events are mutually exclusive
- When events are not mutually exclusive

a) **When events are mutually exclusive:**

if A and B are two mutually exclusive events, then the probability that any one of them will occur (either A will occur or B will occur) is equal to the sum of two separate probabilities

Let, P (M) and P (N), be probabilities of A and B events respectively

then P (M or N) = P (M+N) = P (M) + P (N)

or P (M ∪ N) = P (M) + P (N)

Example 1: In a game of cards, where a pack contains 52 cards, 4 categories exist namely spade, club, diamond, and heart. If you are asked to draw a card from this pack, what is the probability that the card drawn belongs to either the spade or club category?

Solution: Here, P (Spade or club) = P (Spades) + P (Club)

Where P (Spade) = $\frac{13}{52} = \frac{1}{4}$ and $P\ (Club) = \frac{13}{52} = \frac{1}{4}$

Therefore, P (Spade or Club) = $\frac{1}{4} + \frac{1}{4} = \frac{1}{2}$

- **When events are not mutually exclusive**

If A and B are two not mutually exclusive events then the probability that any one of two events will occur (at least one of them) is:

P (M+N) = P(M) + P(N) – P (M ∩ N)

P (M ∪ N) = P(M) + P(N) – P (M ∩ N)

Statement: If M, N are two events then the probability of occurrence of at least one of the two events P (M∪ $\boldsymbol{N}$) = $\boldsymbol{P(M) + P(N) - P(M \cap N)}$

Proof: Let us consider that the random experiment is performed in which S' is the sample space of an experiment. Probability is defined as

P (M∪ N) $= \frac{n(M \cup N)}{n(S')}$

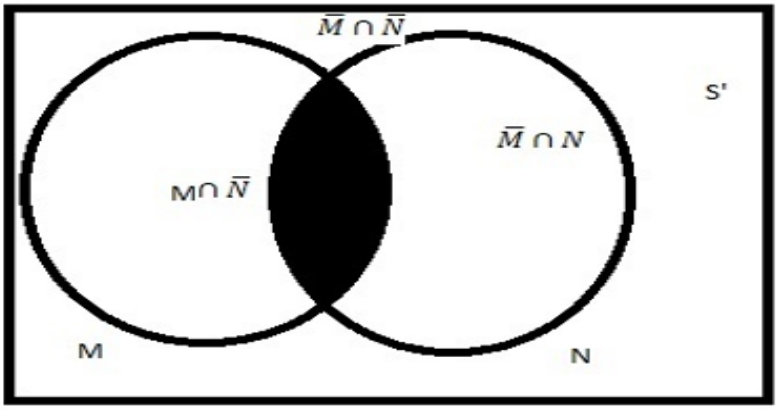
where $n(M \cup N)$ is the number of sample points in favor of event $M \cup N$.

P ($M \cup N$) $= \frac{[n(M)-n(M \cap N)] + [n(M \cap N)+n(N)-n(M \cap N)]}{n(S')}$

$= \frac{n(M)+n(N)-n(M \cap N)}{n(S')}$

$= \frac{n(M)}{n(S')} + \frac{n(N)}{n(S')} - \frac{n(M \cap N)}{n(S')}$

$= P(M) + P(N) - P(M \cap N)$



Remark: For mutually exclusive events Addition theorem of probability reduces to

$$P(M \cap N) = \frac{n(M \cap N)}{n(S')} = \frac{n(\emptyset)}{n(S')} = 0$$

as $M \cap N = \emptyset, n(\emptyset) = 0$.

Example: A card is drawn from a pack of cards. Find the probability that the drawn card is either a black card or either a king.

Sol: Now total black cards = 52

King Cards=4

Black King Cards = 2.

Let P(A)= Probability of black card

P(B)= Probability of King card.

P (A∩B) = Probability of black king card

P(A∪B) =P(A)+P(B)-P(A∩B)

$\frac{26}{52}+\frac{4}{52}-\frac{2}{52}=\frac{28}{52}=\frac{7}{13}$

3.5 MULTIPLICATION THEOREM OF PROBABILITY/ COMPOUND PROBABILITY THEOREM

- When Events are dependent
- When Events are independent

a) When events are dependent:

If A and B are two dependent events then the probability that they occur simultaneously is equal to the product of the probability of event A and the conditional probability of event B on the assumption that A has already occurred. Mathematically, P(MN) = P(M). P (N/M)

or

P (M ∩ N) =P(M). P(N/M)

Proof: Suppose out of total n events, 'm' mutually exclusive equally likes cases, m_1 is favorable event A.

Statement: The probability of simultaneous occurrence of two events M and N is defined as

$P(M \cap N) = P(M).P(N/M), \qquad P(M) \neq 0$

$P(M \cap N) = P(N).P(M/N), \qquad P(N) \neq 0.$

Proof: Let S' represent sample space and M, N are the corresponding events.

$$P(M \cap N) = \frac{n(M \cap N)}{n(S')}$$

$$=\frac{n(M)}{n(S')} \cdot \frac{n(M \cap N)}{n(M)}$$

$$=P(M).P\left(\frac{N}{M}\right)$$

Similarly, $$P(M \cap N) = \frac{n(N)}{n(S')} \cdot \frac{n(M \cap N)}{n(N)}$$

$$=P(N).\,P\left(\frac{M}{N}\right).$$

When events are Independent:

The multiplication theorem states that if A and B are two independent events then the probability, they will occur simultaneously is equal to the product of their independent probabilities.

Mathematically

P (MN) = P(M). P (N)

Or P (M∩ N) =P (M). P (N)

Proof.: For two dependent events we have proved that

P (M∩N) =P(M). P(N/M)

If M and N are two independent events then P(N/M) =P(N) Put in (1)

P (M∩N) =P(M). P(N) Hence Proved.

Example: The probability that A will pass the examination is 0.6. The probability that B will pass the examination is 0.5. Find the probability that at least one of them will pass the examination.

Sol: Let P(A) = Probability that A will pass the examination

P (B)= Probability that B will pass the examination

P(A)=0.6, P(B)=0.5

P($\bar{A}$)=1-P(A)=1-0.6=0.4

P($\bar{B}$)=1-P(B)=1-0.5=0.5

The probability that none will pass =P ($\bar{A}\bar{B}$)

P ($\bar{A}$). P ($\bar{B}$)

=0.4×0.5=0.20

The probability that at least one of them will pass

=1-P (none will pass)

= 1 – 0.20 = 0.80 Ans.

3.6 CONDITIONAL PROBABILITY

Let A and B be two dependent events. When the happening of event A depends upon the happening of event B, then the probability of event A is called conditional probability.

Conditional Probability is denoted by P(A/B).

P(A/B) is the conditional probability of event A, when event B has already occurred.

P(B/A) is the conditional probability of event B, when event A has already occurred.

P(A/B) and P(B/A) are calculated by the following formula:

P(AB) P(A∩B) (a) P(A/B) = P(B) P(B) =

(b) P(B/A) = P(AB) P(A∩B) = P(A) P(A)

For three events A, B, and C.

We know that P(A∩B) = P(B) P(A/B)

P(A∩B) = P(A) P(B/A)

Now P(A∩B∩C) = P(A) P(B/A) P(C/AB)

Here P(A∩B∩C) = Probability of occurrence of A, B and C.

P(B/A) = Probability of B. Given that A has already occurred

P(C/AB) = Probability of C. Given that both A and B already existed

Example: A bag contains 7 red and 5 black balls. Two balls are drawn at random ore other without replacement, Find the probability that both balls drawn are red

Sol: Red= Red balls = 7

Black = Black balls = 5

Total =Red + Black = 7+5=12

Probability of drawing, Red ball in the first attempt, P (R_1) = $\frac{7}{12}$

Probability of drawing, the second Red ball on the condition that the first ball drawn is Red,

$P(R_2/R_1) = \frac{6}{11}$

The probability that both balls drawn are Red is given by

$P(R_1 \text{ and } R_2) = P(R_1R_2) = P(R_1).\ P(R_2/R_1)$

$P(R_1R_2) = \frac{7}{12} \times \frac{6}{11} = \frac{7}{22}$

3.7 GLOSSARY

- **Event**- A collection of one or more outcomes of an experiment.
- **Experiment**- The technique with well-defined outcomes when performed results in one or more than one outcome.
- **Compound event/ Composite event-** An event that contains more than one outcome of a random experiment.
- **Complementary events**- Two events that are taken together include all the outcomes of a random experiment.
- **Classical probability-** It is defined as the method of assigning probabilities to the outcomes or events of an experiment with equal likely possibilities.
- **Conditional probability**- The probability of an event subject to the condition that another event has already occurred.
- **Equally likely outcomes**- Two or more events with the same probability of occurrence.
- **Dependent events-** If the occurrence of one event affects the probability of occurrence of the other event.
- **Independent events**- Two events for which the occurrence of one does not affect the probability of the occurrence of the other.
- **Sample space-** It is defined as collecting all sample points or outcomes of an experiment.

3.8 SUM UP

In the present lesson how one can tackle the difficulty of uncertainty using probability theory has been discussed. Different approaches to study the concept of probability have been focused. Two fundamental laws of probability addition and multiplication have been discussed. The need for conditional probability is emphasized. Difference in different types of events such as mutually exclusive, complementary, and equally likely events is discussed. Different statistical concepts

which help to analyze and interpret real-life data are discussed in detail.

3.9 QUESTIONS FOR PRACTICE

A. Long Answer Type Questions

Q1. State and prove the addition theorem of probability.

Q2. State the conditional theorem of probability and also prove the same.

Q3. Discuss the different approaches to probability.

Q4. There are three events M_1, M_2, M_3 of which must and only one can happen. The odds are 7 to 4 against M_1 and 5 to 3 against M_2. Find the odds against M_3.

Q5. What is the probability that a random leap year will contain 53 Sundays?

Q6. In a class of 25 students with roll no. 1 to 25, a student is picked up a random to answer questions. Find the probability that the roll no of the selected students is either a multiple of 5 or 7.

Q7. The probability of student 'M' passing an examination is 2/9 and of student 'N' passing is 5/9. Assuming the two events 'M' passes and 'N' passes are independent. Find the probability

i) only 'M' passing the examination

ii) only one of them passes the examination.

Q8. A committee of two is selected from two men and two women. Find the probability that the committee contains i) No man ii) one man iii) two men.

B. Short Answer Type Questions

Q1. Given that 'M' and 'N' are two events such that P(M)=0.6, P(N)=0.3 and $P(M \cap N) = 0.2$. Find $P\left(\frac{M}{N}\right)$ and $P\left(\frac{N}{M}\right)$.?

Q2. A bag containing 5 red, 7 green and 4 white balls. Three balls are drawn one after another without replacement. Find the probability that the balls are white and green.

Q3. Suppose that 'n' persons are seated on 'n' chairs at a round table. Find the probability that two specified persons are sitting next to each other.

Q4. A committee of four has to be constituted of 3 economists, 4 engineers, 2 statisticians and 1 doctor. What is the probability that each of the four professionals is represented on the committee?

Q5.Given that the probability a man will be alive 25 years is 0.3 and the probability that his wife will be alive 25 years is 0.4. Find the probability that 25 years hence

i) both will be alive

ii)only the man will be alive

Q8. Two digits are selected at random from the digits 1 to 9. Find the probability that their sum is even.

Q9. M speaks the truth 2 out of 3 times and N 4 out of 5 times, they agree in the response that from a bag containing 6 balls of different colors, a black ball has been drawn. Find the probability that the statement is true.

3.10 MULTIPLE CHOICE QUESTIONS (MCQ)

Q.1. An event in the probability that will never happen is called a:

a) Unsure event b) Sure event c) Possible event d) Impossible event

Q.2. What will be the value of P (not E) if P(E) = 0.07?

a) .90 b) .007 c) .93 d) .72

Q.3. What will be the probability of getting odd numbers if a dice is thrown?

a) 1/2 b) 2 c) 4/2 d) 5/2

Q.4. What is the probability of getting a sum as 3 if a dice is thrown?

a) 2/18 b) 1/18 c) 4 d) 1/36

Q.5. What is the probability of getting an even number when a dice is thrown?

a)1/6 b)1/2 c)1/3 d)1/4

Q.6. The probability of getting two tails when two coins are tossed is -

a)1/6 b)1/2 c) 1/3 d)1/4

Q.7. What is the probability of getting the sum as a prime number if two dice are thrown?

a) 5/24 b) 5/12 c) 5/30 d) 1/4

Q.8. What is the probability of getting at least one head if three unbiased coins are tossed?

a) 7/8 b)1/2 c) 5/8 d) 8/9

Q.9. What is the probability of getting 1 and 5 if a dice is thrown once?

a) 1/6 b)1/3 c) 2/3 d) 8/9

Q.10. What will be the probability of losing a game if the winning probability is 0.3?

a) 0.5 b) 0.6 c) 0.7 d) 0.8

Q.11. Which of these represents the multiplication theorem of probability?

a)$P(A \cap B) = P(B)P(B/A)$ b) $P(A \cap B) = P(A)P(B/A)$

c) $P(A \cap B) = P(B)P(B/B)$ d) $P(A \cap B) = P(A)P(A/A)$

Q.12. A bag contains 5 brown and 7 black pebbles. What is the probability of drawing a brown pebble if the first pebble drawn is black? The balls drawn are not replaced in the box

a) 5/11 b) 8/11 c) 4/18 d) 14/11

Q.13. A bag contains 3 red and 4 blue marbles. Two marbles are drawn without replacement. What is the probability that the second ball is red if it is known the first marble is red?

a) 3/7 b) 4/7 c) 1/3 d) 1/7

Q.14. A bag contains 4 red and 7 blue balls. What is the probability of drawing a blue ball if the first ball drawn is red? The balls drawn are replaced in the bag

a) 8/11 b) 7/11 c) 2/11 d) 3/11

Q.15. Neha has 4 yellow t-shirts, 6 black t-shirts, and 2 blue t-shirts to choose from her outfit today. She chooses a t-shirt randomly with each t-shirt equally likely to be chosen. Find the probability that a black or blue T-shirt is chosen for the outfit.

a) 8/13 b) 5/6 c) 1/2 d) 7/12

Q.16. There are a total of 50 distinct books on a shelf such as 20 math books, 16 physics books, and 14 chemistry books. Find the probability of getting a book that is not a chemistry book or not a physics book.

a) 4/17 b) 43/50 c) 12/31 d) 1

Q.17. A number is selected from the first 20 natural numbers. Find the probability that it would be

divisible by 3 or 7?

a) 19/46 b) 24/67 c) 12/37 d) 7/20

Q.18. There are 24 red marbles in a bag of 68 marbles and 8 of those marbles are both red and white striped. 27 marbles are white striped and of those marbles, the same 8 marbles would be red and white striped. Find the probability of drawing out a marble from the bag that is either red or white striped.

a) 12/35 b) 43/68 c) 26/68 d) 32/55

Q.19. If the spinner has 3 equal sectors colored yellow, blue, and red, then the probability of landing on red or yellow after spinning this spinner is …

a) 2/3 b) 4/7 c) 6/17 d) 23/47

Q.20. In a secondary examination. 75% of the students have passed in History and 65% in Mathematics, while 50% passed in History and Mathematics. If 35 candidates failed in both subjects, what is the total number of candidates sitting for that exam?

a) 658 b) 398 c) 764 d) 350

Answer Key of MCQs:

1. D	2. C	3. A	4. B	5. B
6. D	7. B	8. A	9. B	10. C
11. B	12. A	13. D	14. B	15. D
16. D	17. D	18. B	19. A	20. D

3.12NUMERICAL EXAMPLES

1. **The probability of a student passing in science is ⅘ and of the student passing in both scienceand maths is ½. What is the probability of that student passing in maths knowing that he passedin science?**

Solution: Let A ≡ event of passing in scienceB ≡ event of passing in maths Given, P(B) = ⅘ and P(A ∩ B) = ½

Then, probability of passing maths after passing in science = P(B|A) = P (A ∩ B)/P(A)

= ½ ÷ ⅘ = ⅝

∴ the probability of passing in maths is ⅝.

2. **In a survey among a few people, 60% read Hindi newspapers, 40% read English newspapers and20% read both. If a person is chosen at random and if he already reads an English newspaper find the probability that he also reads a Hindi newspaper.**

Solution: Let there be 100 people in the survey, then the Number of people who read Hindi newspapers = n(A) = 60 Number of people who read English newspapers = n(B) = 40Number of people who read both = n (A ∩ B) = 20

The probability of the person reading a Hindi newspaper when he already reads an English newspaper isgiven by –

P(A|B) = n (A ∩ B)/n(B) = 20/40 = ½.

3. **A fair coin is tossed twice such that E: is the event of having both head and tail and F: is the event ofhaving at most one tail. Find P(E), P(F) and P(E|F)**

Solution: The sample space S = {HH, HT, TH, TT}E = {HT, TH}

F = {HH, HT, TH}E ∩ F = {HT, TH}P(E) = 2/4 = ½ P(F) = ¾

P (E ∩ F) = 2/4 = ½

P(E|F) = P (E ∩ F)/P(F) = ½ ÷ ¾ = ⅔.

4. **In a class, 40% of the students like Mathematics and 25% of students like Physics and 15% like both subjects. One student selects at random, and find the probability that he likes Physics if itis known that he likes Mathematics.**

Solution: Let there be 100 students, then, Number of students like Mathematics = n(A) = 40

Number of students like Physics = n(B) = 25

Number of students like both Mathematics and Physics = n (A ∩ B) = 15

Now, the probability that the student likes Physics if it is known that he likes Mathematics is given by – P(B|A) = n (A ∩ B)/n(A) = 15/40 = ⅜.

5. **Two dice are rolled, if it is known that at least one of the dice always shows 4, find the probability that the numbers that appeared on the dice have a sum of 8.**

Solution: Let, A: one of the outcomes is always 4B: sum of the outcomes is 8

Then, A = {(1, 4), (2, 4), (3, 4), (4, 4), (5, 4), (6, 4), (4, 1), (4, 2), (4, 3), (4, 5), (4, 6)}

B = (4, 4), (5, 3), (3, 5), (6, 2), (2, 6)}n(A) = 11, n(B) = 5, n (A ∩ B) = 1 P(B|A) = n(A ∩ B)/n(A) = 1/11.

6. **A bag contains 3 red and 7 black balls. Two balls are drawn at random without replacement. Ifthe second ball is red, what is the probability that the first ball is also red?**

Solution: Let A: event of selecting a red ball in first drawB: event of selecting a red ball in second draw P (A ∩ B) = P (selecting both red balls) = 3/10 × 2/9 = 1/15

P(B) = P (selecting a red ball in the second draw) = P (red ball and red ball or black ball and red ball)

= P (red ball and red ball) + P (black ball and red ball)

= 3/10 × 2/9 + 7/10 × 3/9 = 3/10

∴ P(A|B) = P (A ∩ B)/P(B) = 1/15 ÷ 3/10 = 2/9.

7. **If a family has two children, what is the conditional probability that both are girls if there is atleast one girl?**

Solution: Let A: both being girlsB: At least one girl n(A) = 1

n(B) = 3

n (A ∩ B) = 1

P(A|B) = n (A ∩ B)/n(B) = ⅓.

8. **A dice and a coin are tossed simultaneously. Find the probability of obtaining a 6, given that ahead appeared.**

Solution: Let A: six coming with a head B: coin shows a head

A = {(6, H))

B = {(1, H), (2, H), (3, H), (4, H), (5, H), (6, H)}

N (A ∩ B) = 1 and n(B) = 6

∴ The probability of getting a six when there is a head is given by –P(A|B) = n (A ∩ B)/n(B) = ⅙.

9. A coin is tossed, and a die is rolled. What is the probability that the coin shows the head andthe die shows 3?

Solution: When a coin is tossed, the outcome is either a head or a tail. Similarly, when a die is rolled, theoutcomes will be 1, 2, 3, 4, 5, 6.

Hence, the required probability = (1/2) (1/6) = 1/12.

3.12 SUGGESTED READINGS

- P.L. Meyer, Introductory probability and statistical applications, Oxford Pub. (1990).
- V.R. Rohatgi and A.K.M.E. Saleh, An Introduction to probability theory and mathematical statistics, Wiley Eastern (2010).
- S.P. Gupta, Statistical Methods, S. Chand and Company, New Delhi.
- A.M. Goon, M.K. Gupta and B. Dasgupta, Fundamental of statistics, World press Calcutta.
- G. Roussas, Introduction to probability, Elsevier, Second Edition (2014).
- P.S. Mann, Introductory Statistics, Wiley India, Seventh edition.

CERTIFICATE/ DIPLOMA IN STATISTICAL ANALYSIS AND RESEARCH METHODOLOGY

SEMESTER II

TIME SERIES ANALYSIS AND PROBABILITY DISTRIBUTIONS

UNIT 4: PROBABILITY DISTRIBUTION: BINOMIAL DISTRIBUTION

STRUCTURE

4.0 OBJECTIVES

After studying this unit, learners will be able to know:

- Meaning of binomial distribution and its properties

- identify the situations where these distributions are applied
- fitting of the binomial distribution to the given data
- define and explain Poisson distribution;
- know as to how a Poisson distribution is fitted to the observed data.

4.1 INTRODUCTION

A representation of every possible value for a discrete random variable together with the probability that each value will occur. It is called a discrete probability distribution. The discrete probability distribution has two different types of distributions.

i) Poisson distribution
ii) Binomial distribution

Let's get into more detail about these two distributions.

4.2 BINOMIAL DISTRIBUTION

The Binomial Distribution (B.D.) was discovered by Swiss Mathematician James Bernoul (1654-1705) and was first published in 1713, eight years after his death. This Distribution is, therefore, also known as "Bernoulli Distribution". The binomial distribution describes discrete, not continuous, data resulting from an experiment known as Bernoulli Process. A binomial distribution is a probability distribution expressing the probability of one set of dichotomous alternatives, i.e., success or failure. In probability theory and statistics, the binomial distribution is the discrete probability distribution that gives only two possible results in an experiment, either Success or Failure. For example, if we toss a coin, there could be only two possible outcomes: heads or tails, and if any test is taken, then there could be only two results: pass or fail. This distribution is also called a binomial probability distribution. There are two parameters n and p used here in a binomial distribution. The variable ‘n’ states the number of times the experiment runs and the variable ‘p’ tells the probability of any one outcome.

For example, as we already know, binomial distribution gives the possibility of a different set of outcomes. In real life, the concept is used for:

- Finding the quantity of raw and used materials while making a product.
- Taking a survey of positive and negative reviews from the public for any specific product or place.

- By using the YES/ NO survey, we can check whether the number of persons view the particular channel.
- To find the number of male and female employees in an organization.
- The number of votes collected by a candidate in an election is counted based on 0 or 1 probability.

As per this distribution, the probability of getting 0, 1, 2, On heads (or tails) in n tosses of an unbiased coin will be given by the successive terms of the expansion of (q + p) n, where p is the probability of success (heads) and q is the probability of failure (i.e. = 1ñ p). The binomial law of probability distribution is applicable only when:

a) A trial results in either the success or failure of an event.

b) The probability of success p remains constant in each trial.

c) The trials are mutually independent i.e.; the outcome of any trial is neither affected by others nor affects others.

4.3 ASSUMPTIONS/ CONDITIONS

i) Each trial has only two possible outcomes either Yes or No, success or failure, etc.

ii) Regardless of how many times the experiment is performed, the probability of the outcome, each time, remains the same.

iii) The trials are statistically independent.

iv) The number of trials is known and is 1, 2, 3, 4, 5, etc.

4.4 PROPERTIES BINOMIAL DISTRIBUTION

The properties of the binomial distribution are:

- There are two possible outcomes: true or false, success or failure.
- There is *n* number of independent trials or a fixed number of n times repeated trials.
- The probability of success or failure remains the same for each trial.
- Only the number of successes is calculated out of *n* independent trials.
- The mean of the binomial distribution is *np*.
- Variance of the binomial distribution is *npq* or *np(1-p)*
- Every trial is an independent trial, which means the outcome of one trial does not affect the outcome of another trial.
- Mean is greater than variance.

- SD = (σ)=$\sqrt{npq}$; Skewness = (q-p)/ $\sqrt{npq}$ Kurtosis=(1-6pq) \ $\sqrt{npq}$
- For fixed n,

  The shape of the curve is:

  (a) if p = q, the distribution will be symmetrical.

  (b) if p<q, the distribution will be positively skewed.

  (c) if p>q, the distribution will be negatively skewed.
- The mean of binomial distribution increases if n increases.
- Since binomial distribution deals with the discrete variable, therefore a smooth frequency can't be drawn exactly rather, presented with a bar diagram.
- The shape of Binomial distribution changes as p changes for a given changes for a given p. As p increases for a fixed, n, the binomial distribution shifts to the right.
- As n increases for a fixed p, the binomial distribution moves to the right, flattens and spreads.
- If n is large (i.e. n→∞) and if neither p nor q is too close to zero, the binomial distribution can be closely approximated by a normal distribution with standardized variable given by z= $\frac{x-np}{\sqrt{npq}}$
- If two independent random variables X and Y follow binomial distribution with parameters (n_1, p) and (n_2, p) respectively, their sum X+Y also follows binomial distribution with parameters (n_1, n_2, p).

4.5 BINOMIAL PROBABILITY FORMULA/DEFINITION

A discrete random variable X is said to follow binomial distribution with parameters n and p if it assumes only a finite number of non-negative integer values and its probability mass function is given by

$$P(r) = {}^{n}C_{r}\, p^{r}\, q^{n-r}$$

where, P (r) = Probability of r successes in n trials;

n = the number of experiments

r = 0, 1, 2, 3, 4, …

p = Probability of Success in a single experiment

q = Probability of Failure in a single experiment = 1 – p

The determining equation for ${}^{n}C_{r}$ can easily be written as:

$${}^{n}C_{r} = \frac{n!}{r!(n-r)!}$$

n! can be simplified as follows:

n! = n (n-1)! = n (n-1) (n-2)! = n (n-1) (n-2) (n-3)! and so on.

Hence the following form of the equations, for carrying out computations of the binomial probability is perhaps more convenient.

$$P(r) = \frac{n!}{r!(n-r)!} p^r q^{n-r}$$

Symbol! means factorial, which is computed as follows: 5! means $5 \times 4 \times 3 \times 2 \times 1 = 120$. Mathematicians define 0! as 1

If n is large in number, say, ${}^{50}C_3$, then we can write (with the help of the above explanation)

$${}^{n}C_r = \frac{50!}{3!(50-3)!} = \frac{(50)\,(49)(48)(47)!}{3!(47)!}$$

$$= \frac{(50)\,(49)(48)(47)!}{3!(47)!}$$

$$= \frac{(50)\,(49)(48)}{3 \times 2 \times 1}$$

As similar,

$${}^{75}C_5 = \frac{75!}{5!(75-5)!} = \frac{(75)(74)(73)(72)(71)\,(70)}{5!(70)!}$$

$$= \frac{(75)(74)(73)(72)(71)}{5 \times 4 \times 3 \times 2 \times 1}$$

Remark:

i) The binomial distribution is the probability distribution of sum of n independent Bernoulli variates.

ii) If X is binomially distributed with parameters n and p, then we may write it as X ~ B (n, p).

iii) If X and Y are two binomially distributed independent random variables with parameters (n_1, p) and (n_2, p) respectively then their sum also follows a binomial distribution with parameters $n_1 + n_2$ and p. But, if the probability of success is not same for the two random variables, then this property does not hold.

4.6 APPLICATIONS OF BINOMIAL DISTRIBUTION

Here are some common applications of the binomial distribution in various fields:

- Coin tosses
- Roll of dice
- Defective items in a production line
- Survey responses (yes/no questions)
- Success/failure rates in medical trials
- Customer behavior (purchase/no purchase)

4.7 CHARACTERISTICS/ FEATURES OF A BINOMIAL DISTRIBUTION

i) The form of the distribution depends upon the parameters *p* and *n*.

ii) The probability that there are *r* successes in *n* no. of trials is given by

$$P(r) = {}^{n}C_r\, p^r\, q^{n-r} = \frac{n!}{r!(n-r)!}\, p^r\, q^{n-r}$$

iii) It is mainly applied when the population being sampled is infinite.

iv) It can also be applied to a finite population if it is not very small or the units sampled are replaced before the next trial is attempted. The point worth noting is *p* should remain unchanged

Example 1: A fair coin is tossed six times. What is the probability of obtaining four or more heads?

Solution: When a fair coin is tossed, the probabilities of head and tail in the case of an unbiased coin are equal, i.e., p = q = ½ or 0.5

∴ The probability of obtaining 4 heads is:

$$P(r) = {}^{n}C_r\, p^r\, q^{n-r} = \frac{n!}{r!(n-r)!}\, p^r\, q^{n-r}$$

$$P(4) = {}^{6}C_4\, (1/2)^4\, (1/2)^{6-4}$$

$$P(4) = \frac{6\times5\times4\times3\times2\times1}{(4\times3\times2\times1)(2\times1)}\, (0.625)\, (0.25)$$

$$= \frac{720}{48}\, (0.625)\, (0.25) = 15 \times (0.625)\, (0.25)$$

$$= 0.234$$

Example 2: If a coin is tossed 5 times, 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

$$P(x=2) = {}^{5}C_2\, p^2\, q^{5-2} = 5!\,/\,2!\;3! \times \left(\frac{1}{2}\right)^2 \times \left(\frac{1}{2}\right)^3$$

$$= \frac{5*4}{1*2} \times \left(\frac{1}{2}\right)^2 \times \left(\frac{1}{2}\right)^3 \qquad = 10*\left(\frac{1}{4}\right)*\left(\frac{1}{8}\right)$$

$$P(x=2) = \frac{5}{16}$$

(b) For at least four heads,

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

Hence, $P(x = 4) = {}^5C_4\, p^4\, q^{5-4} = \frac{5}{1} \times (\frac{1}{2})^4 \times (\frac{1}{2})^1 = \frac{5}{32}$

$P(x = 5) = {}^5C_5\, p^5\, q^{5-5} = (\frac{1}{2})5 = \frac{1}{32}$

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

Example 3: In a game of darts suppose you have a 25% chance that you will hit the bullseye. If you take a total of 15 shots then what is the probability that you will hit the bullseye 5 times?

Solution: $n = 15$, $p = 25 / 100 = 0.25$, $x = 5$

We have to use the Binomial probability distribution given by

$P(X = x) = {}^nC_x\, p^x(1-p)^{n-x}$

$P(X = 5) = ({}^{15}C_5)0.25^5(1-0.25)^{15-5} = 0.165$

CHECK YOUR PROGRESS

Q1.What is meant by binomial distribution?

Ans:_____

_____

Q2.Give formula for the binomial distribution.

Ans:_____

_____

Q3.What are the criteria for the binomial distribution?

Ans:_____

_____

Q4.A fair coin is tossed 10 times, what are the probability of getting exactly 6 heads and at least six heads.

Q5.An unbiased coin is tossed six times.

Find the probability of obtaining (i) exactly 3 heads (ii) less than 3 heads (iii) more than 3 heads (iv) at most 3 heads (v) at least 3 heads (vi) more than 6 heads

4.8 MEAN AND VARIANCE FOR THE BINOMIAL DISTRIBUTION

As discussed in the Introduction, the binomial distribution has expected values of mean (μ) and a standard deviation (σ). We now see the computation of both these statistical measures.

We can represent the mean of the binomial distribution as:

Mean (μ) = np.

where, n = Number of trials; p = probability of success

And, we can calculate the variance by:

Variance, $\sigma^2 = npq$

And standard deviation by:

$\sigma = \sqrt{npq}$

where, n = Number of trials; p = probability of success; and q = probability of failure = 1-p

Example 4: If the probability of defective bolts is 0.1, find the mean and standard deviation for the distribution of defective bolts in a total of 50.

Solution: P = 0.1, n = 500

∴ Hence (μ) = np = 500 × 0.1 = 50

Thus, we can expect 50 bolts to be defective.

Standard Deviation (σ) = npq

n = 500, p = 0.1, q = 1 - p = 1 - 0.1 = 0.9

∴ $\sigma = \sqrt{500 \times .1 \times .9}$

= 6.71

Example 5: In a Binomial Distribution the mean and variance are 85 and 95 respectively. Comment it

Solution: Given mean =np=80 ……(1)

and variance = npq = 90……(2)

Putting the value of np from equation (1) in equation (2) we get

85 (q) = 95

Q= 95/85 >1

The probability can never be more than 1. And mean is greater than variance is one of the properties of binomial distribution. Therefore, the statement is not relating to binomial distribution.

Example 6: In a Binomial Distribution the mean and variance are 85 and 67 respectively. Comment it

Solution: Given mean =np=85 ……(1)

and variance = npq = 67……(2)

Putting the value of np from equation (1) in equation (2) we get

85 (q) = 67

q = 67/85 <1

The probability can never be more than 1 and here mean is greater than variance. Therefore, the statement relates to binomial distribution.

Example 7: In a Binomial Distribution the mean and standard deviation are 80 and 8 respective Find Binomial Distribution.

Solution: Given Mean =80 _____(A)

and Standard Deviation =8

Therefore, np= 80 and $\sqrt{npq} = 8$,

Squaring on both sides, we get

npq = 64 _____(B)

Putting the value of np from equation (A) in equation (B) we get,

80q=64

q = 0.8

p=1-q

1-0.8=0.2

Putting the value of P in equation (A) we get

n (0.2) =80

n= 80/0.2 = 400

Required Binomial Distribution is=$(q+p)^n = (0.8+0.2)^{400}$

4.9 FITTING OF BINOMIAL DISTRIBUTION

When a binomial distribution is to be fitted to observed data, the following procedure is adopted:

i) Determine the values of p and q. If one of these values is known, the other can be found out by the simple relationship p = 1-q and q = 1-p. here, there are three situations:
   - If p and q are equal, we can say, the distribution is symmetrical.
   - If p and q are not equal, the distribution is skewed.
   - The distribution is positively skewed, in case p is less than 0.5, otherwise, it is negatively skewed.

ii) Expand the binomial $(p + q)^{n.}$ The power n is equal to one less than the number of terms in the expanded binomial. For example, if 4 coins are tossed (n = 4) there will be five terms, when 7 coins are tossed (n = 7) there will be 8 terms, and so on.

iii) Multiply each term of the expanded binomial by N (the total frequency), in order to obtain the expected frequency in each category.

If an experiment of n Bernoulli trails is repeated N number of times. Then expected frequency of r successes is given by the formula:

$N * P(r) = N * ({}^{n}C_{r})\, q^{n-r}\, p^{r}$

$r = 0,1,2\ldots\ldots n$

Example 8: Fit the binomial distribution to the following set of data

X	F	FX
0	28	0
1	62	62
2	46	92
3	10	30
4	04	16
	$\sum f = 150$	$\sum fx = 200$

Solution: Here n= 5, $N=\sum f = 150$, Then we find the mean of the distribution as follows,

$\bar{X} = = \sum fx/\sum f = (0 + 62 + 92 + 30 + 16)/150 = 200/150 = 4/3.$

$\bar{X} = np \Longrightarrow 4/3 = 4p \Longrightarrow p=1/3.$

The binomial probabilities are given by,

$P(x) = ({}^{n}C_{r})\, q^{n-r}\, p^{r}$

The expected frequencies are given by the formula,

$N \times P(x) = N \times ({}^{n}C_{r})\, q^{n-r}\, p^{r}$

Substituting r=0, 1, 2, 3, 4, 5 in the above formula we obtain the below,

Fitting of Binomial Distribution:

X	P (x)	Expected Binomial Frequency F(x)= N.p(x) = 150. p(x)
0	${}^{4}C_{0} (\frac{1}{3})^{0} (\frac{2}{3})^{4} = \frac{16}{81} = 0.1975$	$150(0.1975) \times 29.63 \cong 30$
1	${}^{4}C_{1} (\frac{1}{3})^{1} (\frac{2}{3})^{3} = \frac{4*8}{81} = 0.3951$	$150(0.3951) \times 59.26 \cong 59$
2	${}^{4}C_{2} (\frac{1}{3})^{2} (\frac{2}{3})^{2} = \frac{4*3}{1*2} \times \frac{4}{81} = 0.2963$	$150(0.2963) \times 44.44 \cong 44$
3	${}^{4}C_{3} (\frac{1}{3})^{3} (\frac{2}{3})^{1} = \frac{4*2}{81} = 0.0988$	$150(0.0988) \times 14.81 \cong 15$

4	${}^4C_4 \left(\frac{1}{3}\right)^4 \left(\frac{2}{3}\right)^0 = \frac{1}{81} = 0.0123$	$150(0.0123) \times 1.85 \cong 2$

4.10 POISSON DISTRIBUTION

Poisson distribution, developed by a French mathematician Simeon Poisson, is so known by his name. It deals with counting the number of occurrences of a particular event in a specific time interval or region of space. It is used in practice where there are infrequently occurring events for time, volume (similar units), area, etc.

Poisson distribution is a discrete probability function, the variable can only take specific values from a specified set of numbers. The Poisson distribution calculates the probability that an event will occur during a given time interval, denoted by "x". Stated differently, we can characterize it as the probability distribution that emerges from the Poisson task. A Poisson experiment is a statistical test that divides the data into two groups, such as successful and unsuccessful. A limiting process of the binomial distribution is the Poisson distribution. The Poisson distribution process corresponds to a Bernoulli process with a very large number of trials (n) and a very low probability of success.

For example: Our interest may lie in how many printing mistakes are there on each page of a book but we are not interested in counting the number of words without any printing mistake. Second example is as in production where control of quality is the major concern, it often requires counting the number of defects (and not the non-defects) per item.

Poisson distribution is a limiting case of binomial distribution under the following conditions:

i) n, the number of trials is indefinitely large, i.e. $n \to \infty$.

ii) p, the constant probability of success for each trial is very small, i.e. $p \to 0$.

iii) np is a finite quantity say 'λ'

A Poisson random variable "x" defines the number of successes in the experiment. This distribution occurs when there are events that do not occur as the outcomes of a definite number of outcomes. Poisson distribution is used under certain conditions. They are:

- The number of trials "n" tends to infinity
- Probability of success "p" tends to zero
- np = 1 is finite

4.11 FORMULA / DEFINITION OF POISSON DISTRIBUTION

A random variable X is said to follow Poisson distribution if it assumes an indefinite number of non-negative integer values and its probability mass function is given below. This would comparatively be simpler to deal with and is given by the Poisson distribution formula as follows:

$$P(r) = \frac{e^{-m}\, m^r}{r!}$$

where, p (r) = Probability of successes

r = 0, 1, 2, 3, 4,…. ∞ (any positive integer)

e = a constant with value: 2.7183 (the base of natural logarithms)

m = The mean of the Poisson Distribution, i.e., np or the average number of occurrences of an event.

Remark

i) If X follows Poisson distribution with parameter m, then we shall use the notation X ~ P(m).

ii) If X and Y are two independent Poisson variates with parameters m_1 and m_2 respectively, then X + Y is also a Poisson variate with parameter m_1+m_2. This known as additive property of Poisson distribution.

4.12 CHARACTERISTICS OF THE POISSON DISTRIBUTION

- It is also a discrete probability distribution and it is the limiting form of the binomial distribution.
- The range of the random variable is $0 \leq r < \infty$
- It consists of a single parameter *m* only. So, the entire distribution can be obtained by knowing this value only.
- It is a positively skewed distribution. The skewness, therefore, decreases when *m* increases.

4.13 PROPERTIES OF POISSON DISTRIBUTION

The Poisson distribution is a probability distribution that describes the number of events that occur within a fixed interval of time or space, given a known average rate of occurrence. Here are some key properties of the Poisson distribution:

1) Poisson Distribution is a discrete distribution where a number of successes are in whole numbers such as 0, 1, 2, ...and not in fractions.
2) The value of p is very small and close to zero [generally < 0.1] and q is very high close to 1; p+q = 1
3) n is very large and approaching infinity but is finite.

4) Poisson Distribution is a skewed distribution. If n is constant and m increases the distribution shifts to the right and skewness is reduced.
5) Poisson distribution can be approximated by the binomial distribution when the number of trials is large, and the probability of success is small.
6) In Poisson Distribution Mean = Variance =m
    - i.e. The mean (μ) of a Poisson distribution is equal to the average rate m =μ=m.
    - The variance (σ^2) is also equal to m. $=\Sigma^2=M$.

4.14 APPLICATIONS OF POSSION DISTRIBUTION

The Poisson distribution has various applications in different fields due to its ability to model the probability of a given number of events occurring in a fixed interval of time or space, given a known average rate of occurrence. Some common applications include:

- Poisson distribution is often used to model the number of cars passing through a traffic checkpoint in a given time, assuming a constant average rate.
- It is used to model the number of calls received by a call center within a specific time frame, assuming calls arrive independently and at a constant average rate.
- Poisson processes are frequently applied to model the arrival of customers in queues, such as those in banks, supermarkets, or service centers.
- The Poisson distribution is employed in insurance to model the number of claims filed within a certain time period, assuming a constant average rate of claims.
- It is used in economic models to represent the occurrence of rare events, such as financial market crashes or defaults.
- Poisson distribution is utilized in biology and medicine to model the distribution of rare mutations or the number of occurrences of specific medical events within a population.
- In particle physics, the distribution of certain types of particles in a detector can be modeled using the Poisson distribution.
- The distribution is applied to model the arrival of packets in a network or the occurrence of events in internet traffic analysis.
- Poisson distribution is used in quality control to model the number of defects in a product or

the number of errors in a process.

- It can be applied to model the occurrence of rare environmental events, such as earthquakes, floods, or ecological disturbances.
- In inventory management, the Poisson distribution may be used to model the demand for a product over time.

Poisson Distribution Table:

	λ									
x	.1	.2	.3	.4	.5	.6	.7	.8	.9	1.0
0	.9048	.8187	.7408	.6703	.6065	.5488	.4966	.4493	.4066	.3679
1	.0905	.1637	.2222	.2681	.3033	.3293	.3476	.3595	.3659	.3679
2	.0045	.0164	.0333	.0536	.0758	.0988	.1217	.1438	.1647	.1839
3	.0002	.0011	.0033	.0072	.0126	.0198	.0284	.0383	.0494	.0613
4	.0000	.0001	.0003	.0007	.0016	.0030	.0050	.0077	.0111	.0153
5	.0000	.0000	.0000	.0001	.0002	.0004	.0007	.0012	.0020	.0031
6	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0002	.0003	.0005
7	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001

	λ									
x	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2.0
0	.3329	.3012	.2725	.2466	.2231	.2019	.1827	.1653	.1496	.1353
1	.3662	.3614	.3543	.3452	.3347	.3230	.3106	.2975	.2842	.2707
2	.2014	.2169	.2303	.2417	.2510	.2584	.2640	.2678	.2700	.2707
3	.0738	.0867	.0998	.1128	.1255	.1378	.1496	.1607	.1710	.1804
4	.0203	.0260	.0324	.0395	.0471	.0551	.0636	.0723	.0812	.0902
5	.0045	.0062	.0084	.0111	.0141	.0176	.0216	.0260	.0309	.0361
6	.0008	.0012	.0018	.0026	.0035	.0047	.0061	.0078	.0098	.0120
7	.0001	.0002	.0003	.0005	.0008	.0011	.0015	.0020	.0027	.0034
8	.0000	.0000	.0001	.0001	.0001	.0002	.0003	.0005	.0006	.0009
9	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0001	.0001	.0002

4.15 MEASURES OF MEAN AND VARIANCE FOR POISSON DISTRIBUTION

In the Poisson distribution, the mean (m) and the variance (s_2) represent the same value, i.e.,

Mean = variance = np = m

Variance = np;

S.D. (σ) = $\sqrt{np}$

Example 9: If 5% of tubes manufactured by a company are defective. Find the chance that out of 100 such tubes produced.

(i) None of the tube is defective

(ii) More than 3 tubes are defective.

Solution: Let X be a random variable representing no. of defective tubes produced. As n is large and p is small, we can use Poisson distribution.

Given, p=0.05

n = 100

m=np=100×0.05=5

The probability of x defective bulbs is given by Poisson probability law as:

$P(r) = \frac{e^{-m} m^r}{r!}$

$= \frac{e^{-5}5^x}{x!}$

a) None of the tube is defective

$P(0) = \frac{e^{-5}5^0}{0!} = e^{-5}$

P (0) = 0.00638

b) More than 3 tubes are defective

P (X>3) = 1-P (X ≤ 3)

=1- [P (0) + P (1) + P (2) + P (3)]

$= 1- (\frac{e^{-5}5^0}{0!} + \frac{e^{-5}1}{1!} + \frac{e^{-5}5^2}{2!} + \frac{e^{-5}5^3}{3!})$

$=1- e^{-5}(1 + 5 + \frac{25}{5} + \frac{125}{6})$

=1- 0.00638 (39.33)

=1- 0.2509

= 0.7491

Example 10: If the probability that an individual suffers a bad reaction from an injection of a given serum is 0.001, determine the probability that out of 500 individuals

i) exactly 3,

ii) more than 2

individuals suffer from bad reaction

Solution: Let X be the Poisson variate, "Number of individuals suffering from bad reaction". Then,

n = 1500, p = 0.001,

m = np = (1500) (0.001) = 1.5

therefore, By Poisson distribution,

$P(x) = \frac{e^{-m}m^X}{X!}$; (x=0,1,2,...)

$= \frac{e^{-1.5}1.5^X}{X!}$

i) The desired probability = P[X = 3]

$= \frac{e^{-1.5}1.5^3}{3!}$

$= \frac{(0.2231)(3.375)}{6}$

= 0.1255

(from table: $e^{-1.5} = e^{-1} \times e^{-0.5}$ = (0.3679) (0.6065) = 0.2231)

ii) The desired probability P (X > 2)

= 1 – [P (X= 2) + P (X= 1) + P (X= 0)]

$= 1 - [\frac{e^{-1.5}1.5^2}{2!} + \frac{e^{-1.5}1.5^1}{1!} + \frac{e^{-1.5}1.5^0}{0!}]$

$= 1- e^{-1.5} [\frac{2.25}{2} + 1.5 + 1]$

$= 1 - e^{-1.5}$ (3.625)

= 1- (0.2231) (3.625)

= 1- 0.8087

= 0.1913

Example 11: **2% of the electronic cars produced in a certain manufacturing process turn out to be defective. What is the probability that a shipment of 200 cars will contain exactly 5 defectives? Also, find the mean and standard deviation.**

Solution: In the given illustration n = 200;

Probability of a defective car (P) = 2/100 = 0.02

Since, n is large and p is small, the Poisson distribution is applicable. Apply the formula

$$P(r) = \frac{e^{-m} m^r}{r!}$$

The probability of 5 defective cars in 200 cars is given by

$$P(5) = \frac{(m^5)e^{-m}}{5!}$$

where m = np = 200 × 0.02 = 4;

e = 2.7183 (constant)

$$P(5) = \frac{45\ \ 2.7183-4}{5*4*3*2*1} = \frac{(1024)\frac{1}{2.7183^4}}{120}$$

$$= \frac{1024*0.0183}{120} = 0.156$$

Mean = np = 200 × 0.02 = 4; σ = np = 4 = 2

CHECK YOUR PROGRESS -B

Q1: What is a Poisson distribution?

Ans:_____

_____

Q2: Give any two properties of Poisson distribution.

Ans:_____

_____

Q3. What are the applications of Poisson distribution?

Ans:_____

_____

4.16 FITTING OF A POISSON DISTRIBUTION

To fit a Poisson distribution to a given observed data (frequency distribution), the procedure is as follows:

1) We must obtain the value of its mean i.e.,

$$\bar{X} = \frac{\Sigma fx}{\Sigma f} = m$$

m = np

2) The probabilities of various values of the random variables (r) are to be computed by using p.m.f. i.e.,

$$P(r) = \frac{e^{-m} m^r}{r!}$$

r = 0,1,2……

3) To get expected frequencies

f (x) = N. p(x)

Example 12: The following table gives the number of days in a 50 days period during automobile accidents occurred in a city:

No. of accidents (X)	No. of Days (f)	fX	N. p(x)
0	21	0	50 × 0.4066 = 20.33
1	18	18	50 × 0.36594 = 18.297
2	7	14	50 × 0.1646 = 8.2337
3	3	9	50 × 0.0490 = 2.47
4	1	4	50 × 0.0111154 = 0.56
	N= $\sum f$ = 50	$\sum fX$ = 45	

$\bar{X} = \frac{\sum fx}{\sum f} = \frac{45}{50} = 0.9$

$\therefore$ m = 0.9

Now frequency of x accidents is as:

$P(x) = \frac{e^{-m} m^x}{x!}$

$P(x) = \frac{e^{-0.9} (0.9)^x}{x!}$

Now,

$P(0) = \frac{e^{-0.9} (0.9)^0}{0!} = e^{-0.9} = 0.4066$

$P(1) = \frac{e^{-0.9} (0.9)^1}{1!} = 0.4066 \times 0.9 = 0.36594$

$P(2) = \frac{e^{-0.9} (0.9)^2}{2!} = (0.4066 \times (0.9)^2) / 2 = 0.1646$

$P(3) = \frac{e^{-0.9} (0.9)^3}{3!} = (0.4066 \times (0.9)^3) / 6 = 0.0490$

$P(4) = \frac{e^{-0.9} (0.9)^4}{4!} = (0.4066 \times (0.9)^4) / 24 = 0.0111154$

4.17 SUM UP

In this unit, we learn about binomial and Poisson distribution, as binomial distribution is the probability distribution that is discrete and applicable to events having only two possible results in an experiment, either success or failure. A few circumstances where we have binomial experiments are tossing a coin: head or tail, the result of a test: pass or fail, selected in an interview: yes/ no, or nature of the product: defective/non-defective. Such a distribution of a binomial random variable

is called a binomial probability distribution. A Poisson experiment is a statistical experiment that classifies the experiment into two categories, such as success or failure. Poisson distribution is a limiting process of the binomial distribution. A Poisson random variable *x* defines the number of successes in the experiment. This distribution occurs when there are events that do not occur as the outcomes of a definite number of outcomes.

4.18 SUGGESTED READINGS

- P.L. Meyer, Introductory probability and statistical applications, Oxford pub (1990).
- V.R. Rohatgi and A.K.M.E. Saleh, An Introduction to probability theory and mathematical statistics, Wiley Eastern (2010).
- S.P. Gupta, Statistical methods, S. Chand and company, New Delhi.
- A.M. Goon, M.K. Gupta and B. Dasgupta, Fundamental of statistics, World press Calcutta.
- G. Roussas, Introduction to probability, Elsevier, Second Edition (2014).
- P.S. Mann, Introductory Statistics, Wiley India, Seventh edition.

CERTIFICATE/ DIPLOMA IN STATISTICAL ANALYSIS AND RESEARCH METHODOLOGY

SEMESTER II

TIME SERIES ANALYSIS AND PROBABILITY DISTRIBUTIONS

UNIT 5: NORMAL DISTRIBUTION- MEANING, PROPERTIES AND FITTING

STRUCTURE

5.0 OBJECTIVES

After reading this unit, you will be able to know:

- the concept of normal distribution

- theoretical basis of the normal probability curve
- Characteristics of the normal probability curve and normal distribution
- Analyse the properties of the normal distribution
- Apply the normal distribution.

5.1 INTRODUCTION

In the middle of the 19th Century Quetelet promoted the applicability of the normal curve. He believes that the normal cure could be extended to apply to problems of anthropology sociology and human affairs. In the latter part of the 19th century, Sir Francis Galton began the first serious study of individual differences and during his systematic study, he found that most of the physical and psychological traits of human beings conformed reasonably well to the normal curve. In this way, he extended the applicability of the normal curve. The normal curve is also known as the Gaussian Curve and bell-shaped curve. A normal curve is one which graphically represents a normal distribution. A normal distribution is one in which the majority of the cases fall in the middle of the scale and a small number of cases are located at both extremes of the scale.

5.2 NORMAL DISTRIBUTION: MEANING

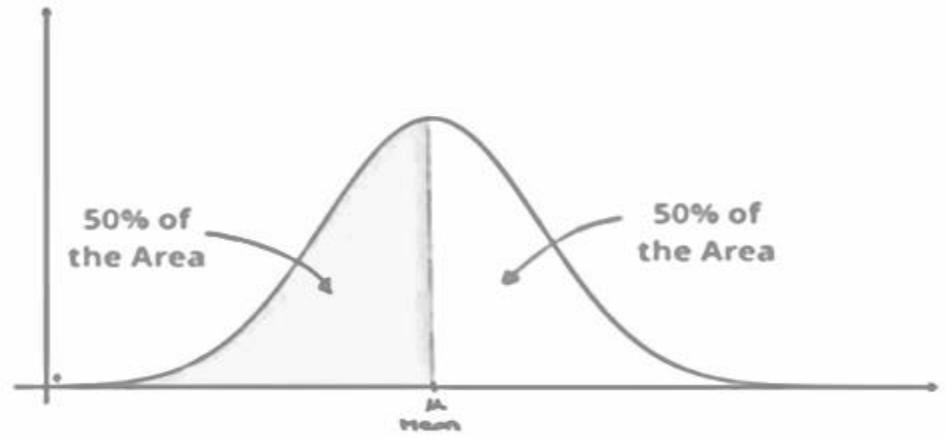
A normal distribution is a fundamental concept in statistics representing a specific pattern of data distribution. Also known as a Gaussian distribution or bell curve, it is characterized by a symmetrical, bell-shaped probability density function. In this distribution, the mean, median, and mode are all centrally located, and the curve is symmetrically distributed on either side of the mean. The normal distribution is defined by two parameters: the mean (μ), which is located at the center of the curve, and the standard deviation (σ), which measures the spread of the distribution. The majority of the data falls within one, two, and three standard deviations from the mean.

The normal distribution is a crucial concept in statistical inference. Many statistical methods, such as hypothesis testing and confidence intervals, assume normality. The Central Limit Theorem further underscores its importance, stating that the sampling distribution of the sample mean tends to be approximately normally distributed, regardless of the original distribution of the population. The normal distribution's ubiquity in nature and its mathematical properties make it a powerful and widely used tool in statistical analysis, providing a framework for

understanding and interpreting various phenomena in diverse fields.

5.3 NORMAL DISTRIBUTION CURVE

Normal probability distribution is determined by two parameters, mean and variance. The normal distributions are used to represent the real-valued random variables whose distributions are unknown. They are used very frequently in the areas of natural sciences and social sciences. When the normal distribution is represented in the form of a graph, it is known as a normal probability distribution curve or simply a normal curve. A normal curve is a bell-shaped curve, bilaterally symmetrical, and is a continuous frequency distribution curve. Such a curve is formed as a result of plotting frequencies of scores of a continuous variable in a large sample. The curve is known as the normal probability distribution curve because its y ordinates provide relative frequencies or the probabilities instead of the observed frequencies. A continuous random variable can be said to be normally distributed if the histogram of its relative frequency has the shape of a normal curve



It is most important to understand the characteristics of the frequency distribution of normal curves in the fields of mental measurement and experimental psychology.

5.4 IMPORTANCE OF NORMAL DISTRIBUTION

In the social sciences as well as the natural sciences, the normal distribution is highly important.

The following lists some of the normal distribution's significance:

- A continuous, normal distribution is important to statistical theory and inference.
- It is a useful method of sampling distribution.

- The normal distribution is a helpful way of sampling distribution because it has several mathematical qualities that make it straightforward to define the frequency distribution in the simplest form.
- The distributions of many behavioral science variables, such as height, weight, achievement, and IQ, roughly resemble the normal curve.
- The normal distribution is a prerequisite for many inferential statistics, including the z-test, t-test, and F-test.

5.5 PROPERTIES OF NORMAL DISTRIBUTION CURVE

The following are the properties of the normal curve:

1. The total area under the curve is one. Each has an area equal to 0.5.
2. The curve is bell-shaped, has a continuous frequency distribution curve, and is bilaterally symmetrical.
3. The normal is bell-shaped and is symmetric about the line $X=\mu$. It has the same shape on either side of the line $X=\mu$.
4. The points of inflexion of the curve are $\mu \pm \sigma$.
5. For a random variable, it is a continuous probability distribution.
6. The normal curve is unimodal i.e., it has only one mode.
7. Its two sides, the right and left, have values of mean, median, and mode that are equal (mean = median = mode), meaning that they coincide at the same point in the middle of the curve.
8. The normal curve is asymptotic, meaning that as it gets farther from the mean, it approaches the x-axis but never touches it.
9. The curve is divided into two halves by the mean, which lies in the middle of the curve. The entire area of the normal curve lies between the mean and $z \pm 3\ \sigma$.
10. It is stated that the mean is zero ($\mu=0$), the variance is one ($\sigma^2=1$), the standard deviation is one ($\sigma = 1$), and the area under the normal curve is one ($N=1$).
11. The term "inflection points" refers to the point where the curve shifts from curving upward to downward.
12. The z-scores or the standard scores in the probability curve towards the right from the mean are positive and towards the left from the mean are negative.
13. The normal distribution is free from skewness, that is, its coefficient of skewness amounts to

zero.

14. The fractional areas in between any two-given z-scores are identical in both halves of the normal curve, for example, the fractional area between the z-scores of +1 is identical to the z-scores of –1. The height of the ordinates at a particular z-score in both halves of the normal curve is the same, for example, the height of an ordinate at +1z is equal to the height of an ordinate at –1z.

5.6 APPLICATION OF THE NORMAL CURVE

The normal curve has the main practical application given below:

- A normal curve helps in transforming the raw scores into standard scores
- To determine the percentage of cases that are above or below a given score or reference point
- To determine the limits of scores which include a given percentage of cases to determine the percentile rank of an individual or a student in his group
- To find out the percentile value of an individual based on his percentile rank
- To compare the two distributions in terms of overlapping
- With the help of a normal curve, we can calculate the percentile rank of the given scores.
- A normal curve is used to find the limits in any normal distribution which includes a given percentage of the cases.
- We can compare two distributions in terms of overlapping with the help of a normal curve
- A normal curve is used to determine the relative difficulty of test questions, problems, and other test items
- When the trait is normally distributed normal curve is used to separate a given group into subgroups according to capacity

CHECK YOUR PROGRESS- A

Q1. What is a normal curve?

Ans:_____

_____

Q2. What are the properties of a normal curve?

Ans:_____

Q3. What are the applications of normal distribution

Ans:_____

_____

5.7 FORMULA FOR THE NORMAL DISTRIBUTION

The formula for the normal distribution is;

$$\frac{1}{\sigma\sqrt{2\,\pi}}\; e^{-\frac{1}{2}\,\left(\frac{x-\mu}{\sigma}\right)^2}$$

Where x is the variable, μ is the mean, σ is the standard deviation.

TABLE OF AREAS UNDER THE NORMAL PROBABILITY CURVE

The normal probability curve table is generally limited to the areas under normal curve with N = 1, σ = 1. In case, when the values of N and σ are different from these, the measurements or scores should be converted into sigma scores (also referred to as standard scores or z scores). The process is as follows:

$Z=\frac{X-\mu}{\sigma}$

Where, z = Standard Score μ = Mean of X Scores X = Raw Score, σ = Standard Deviation of X Scores

The table of areas of the normal probability curve is then referred to find out the proportion of area between the mean and the z value. Though the total area under the N.P.C. is 1, for convenience, the total area under the curve is taken to be 10,000 because of the greater ease with which fractional parts of the total area, may be then calculated. The first column of the table, x/σ gives the distance in tenths of σ measured off on the baseline for the normal curve from the mean as origin. In the row, the x/σ distance is given to the second place of the decimal. To find the number of cases in the normal distribution between the mean, and the ordinate erected at a distance of lσ unit from the mean, we go down the x/σ column until 1.0 is reached and in the next column under .00 we take the entry opposite 1.0, namely 3413. This figure means that 3413 cases in 10,000; or 34.13 percent of the entire area of the curve lies between the mean and lσ. Similarly, if we have to find the percentage of the distribution between the mean and 1 .56 σ,

say, we go down the x/σ column to 1.5, then across horizontally to the column headed by .06, and note the entry 44.06. This is the percentage of the total area that lies between the mean and 1.56σ.

POINTS TO BE KEPT IN MIND WHILE CONSULTING THE TABLE OF AREA UNDER THE NORMAL PROBABILITY CURVE

The following points are to be kept in mind to avoid errors while consulting the N.P.C. Table. Every given score or observation must be converted into a standard measure i.e. Z score, by using the following formula:

$$Z = \frac{X - \mu}{\sigma}$$

The mean of the curve is always the reference point, and all the values of areas are given in terms of distances from the mean which is zero. 3) The area in terms of proportion can be converted into percentage, and 4) While consulting the table, absolute values of z should be taken. However, a negative value of z shows that the scores and the area lie below the mean and this fact should be kept in mind while doing further calculations on the area. A positive value of z shows that the score lies above the mean i.e. right side.

5.8 Practical Problems Related to the Normal Probability Curve

a) P (0 < Z < 1.24)

b) P (1.24 < Z < 2.56)

c) P (Z > 1.96)

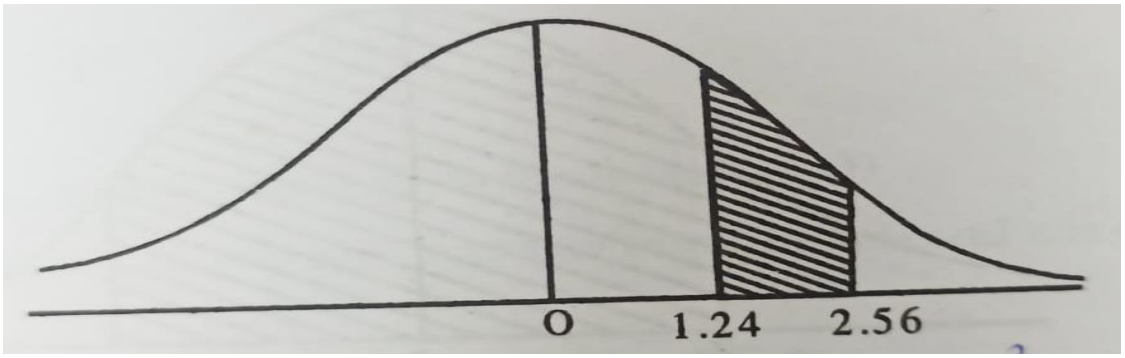
d) P (-2.36 < Z < -1.98)

Solution: (to find with the help of table)

a) P (0 < Z < 1.24) = 0.3925

b) P (1.24 < Z < 2.56)

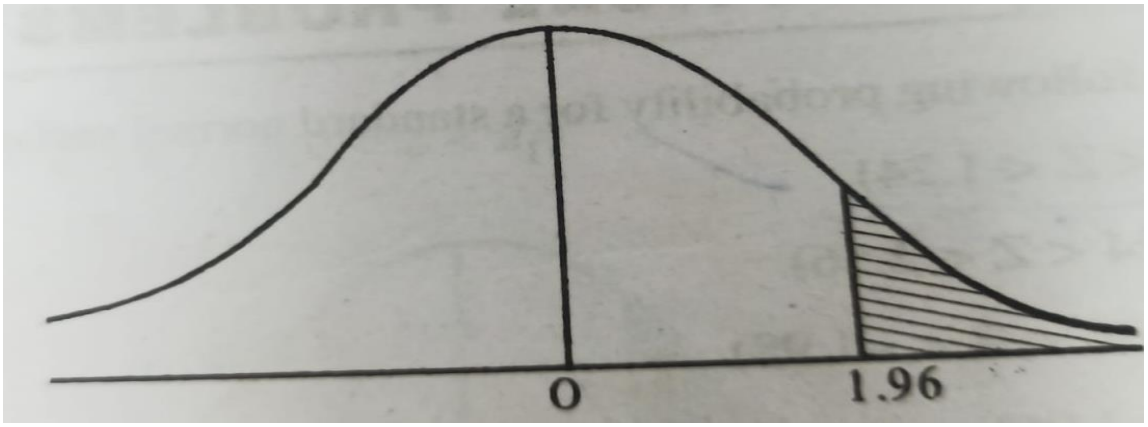
= P (0 < Z < 2.56) - P (0 < Z < 1.24)



= 0.4948 – 0.3925

= 0.1023

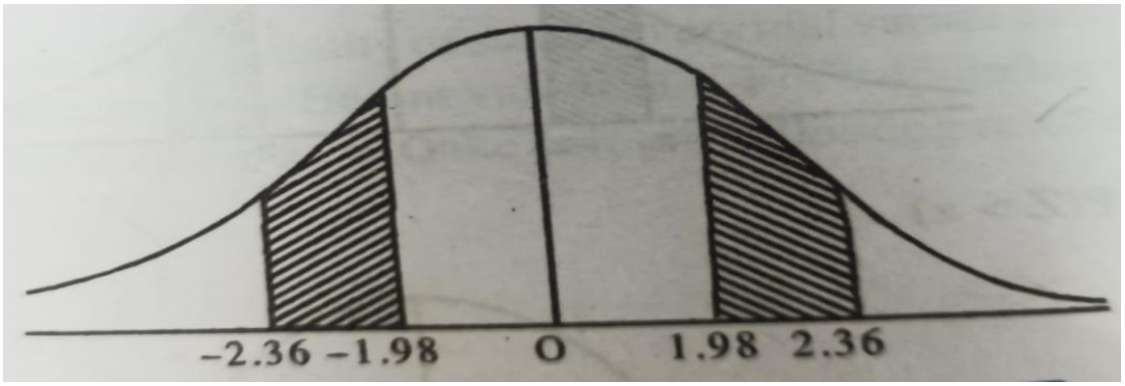
c) P (Z > 1.96)



= 0.5 – P (0 < Z < 1.96)

= 0.025

d) P (-2.36 < Z < -1.98)



= P (-2.36 < Z < -1.98) = P (1.98 < Z < 2.36)

= P (0 < Z < 2.36) - P (0 < Z < 1.98)

= 0.4904 – 0.4761

= 0.0143

Example 1: Assume the mean height of soldiers to be 68.22 inches with a variance of 10.8 How many soldiers in a regiment of 1000 would you expect to be over six feet tall.

Sol: Let X be a random variate representing the height of soldiers, having mean 68.22 variance 10.8.

$\mu = 68.22$

$\sigma^2 = 10.8$

$\sigma = \sqrt{10.8} = 3.286$

The standard normal variate corresponding to X is,

$Z = \frac{X-\mu}{\sigma}$

$= \frac{X-68.22}{3.286}$

We have to find X greater than 6 feet or 72 inches.

when X = 72,

$Z = \frac{X-\mu}{\sigma}$

$= \frac{72-68.22}{3.286}$

$= \frac{3.78}{3.286} = 1.15$

P (X > 72) = P (Z >1.15)

=0.5 – P (0 < Z < 1.15)

= 0.5 - 0.3749 (From table)

= 0.1251

Hence in a regiment of 1000 soldiers the number of soldiers over 6 feet is 1000 × 0.1251=125.1= 125.

Example 2: 1000 tubes with a mean life of 120 days are installed in a new factory: their length of life is normally distributed with standard deviation of 20 days. How many tubes will expire in less than 90 days?

Sol: Let X is a random variable representing the life of bulbs with mean 120 and standard 20.

$\mu=120$

$\sigma = 20$

The standard normal variable corresponding to X is

$Z= \frac{X-\mu}{\sigma}$

$= \frac{X-120}{20}$

When X = 90,

$= \frac{90-120}{20} = -1.5$

So, P (X < 90) = P (Z < -1.5)

= 0.5 - P (-1.5 < Z < 0)

= 0.5 – P (0 < Z < 1.5)

= 0.5 - 0.4332

= 0.0668

Hence, the number of bulbs out of 1000 expected to expire in less than 90 days

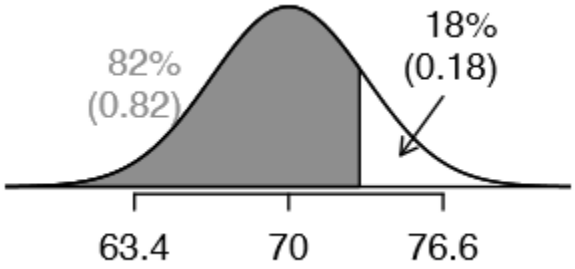
= 1000 × 0.0668

= 66.8 = 67

Example 3: Based on a sample of 100 men, (USDA Food Commodity Intake Database) the heights of male adults between the ages 20 and 62 in the US is nearly normal with a mean 70.0" and a standard deviation 3.3". What is the adult male height at the 82nd percentile?

Sol: we want to find the Z score at the 82nd percentile, which will be a positive value. Looking in the Z table, we find Z falls in row 0:9 and the nearest column is 0.02, i.e. Z = 0.92. Finally, the height x is found using the Z score formula with the known mean μ, standard deviation σ, and Z

score Z = 0.92:



$$Z = \frac{X-\mu}{\sigma} = \frac{X-70}{3.3}$$

This yields 73.04 inches or about 6'1" as the height at the 82nd percentile.

Example 4: For some laptops, the time between charges of the battery is normally distributed with a mean of 50 hours and a standard deviation of 15 hours. Person A has one of these laptops and needs to know the probability that the period will be between 50 and 70 hours.

Solution: Let x be the random variable that represents the time period.

Given Mean, $\mu = 50$

and standard deviation, $\sigma = 15$

To find: Probability that x is between 50 and 70 or P (50 < x < 70)

By using the transformation equation, we know;

$$z = \frac{X-\mu}{\sigma}$$

For x = 50, z = (50 – 50) / 15 = 0

For x = 70, z = (70 – 50) / 15 = 1.33

P (50 < x < 70) = P (0 < z < 1.33) = [area to the left of z = 1.33] – [area to the left of z = 0]

From the table we get the value, such as;

P (0 < z < 1.33) = 0.9082 – 0.5 = 0.4082

The probability that Person A's laptops have a time period between 50 and 70 hours is equal to 0.4082.

Example 5: The mean yield per plot of a crop is 17 kg and standard deviation is 3 kg. If distribution of yield per plot is normal, find the percentage of plots giving yields:

(i) Between 15.5 kg and 20 kg and

(ii) More than 20 kg

Sol: Let X be a random variable representing yield per plot of the crop

Here X is a normal variate with mean 17 and S.D. 3,

$\mu=17, \sigma=3$

The standard normal variate corresponding to X is

$Z=\frac{X-\mu}{\sigma} = \frac{X-17}{3}$

(i) When X=15.5, $Z = \frac{15.5-17}{3} = \frac{-1.5}{3} = -0.5$

When X = 20,

$Z = \frac{20-17}{3} = \frac{3}{3} = 1$

$P(15.5 < X < 20) = P(-0.5 < z < 1)$

$= P(-0.5 < Z < 0) + P(0 < Z < 1)$

$= P(0 < Z < 0.5) + P(0 < Z < 1)$ (Because of symmetry)

$= 0.1915+0.3413$

$= 0.5328$

Thus, percentage of plots giving yield between 15.5 kg and 20 kg

$=0.5328 \times 100 = 53.28\%$

(ii) $P(X > 20) = P(Z > 1)$

$= 0.5 - P(0 < Z < 1)$

$= 0.5 - 0.3413$

$= 0.1587$

Thus, percentage of plots giving yield more than 20 kg $=0.1587 \times 100 = 15.87\%$

CHECK YOUR PROGRESS- B

Q1. From a large group of men, 4% are under 60 inches in height, 40% are between 60 Of enlarges summing a normal distribution, find the mean and standard deviation.

Q2. The monthly salary of 100 workers is normally distributed around a mean of ₹45 and standard deviation ₹5. Estimate the number of workers whose monthly salary women between 44 and 47.

Q3. The net profit of 500 companies is normally distributed with a mean profit of ₹180 lakh and a standard deviation of ₹25 lakhs. Find the number of companies whose profit are:

(i) Less than 158 (ii) More than 190 (iii) Between 110 and 150.

Q4. The mean weight of 100 students in a school is 66.5 kg and standard deviation is 80 kg. Assuming that the weights are normally distributed, how many students' weights:

(i) More than 74.5 kg (ii) Less than 63.3 kg (iii) Between 50.5 kg and 64.1 kg.

Q5. Based on a sample of 100 men, (USDA Food Commodity Intake Database) the heights of male adults between the ages 20 and 62 in the US is nearly normal with mean 70.0" and standard deviation 3.3".

i. What is the probability that a randomly selected male adult is at least 6'2" (74 inches)?
ii. What is the probability that a male adult is shorter than 5'9" (69 inches)?

Answers:

1) 65.42, 3.29,
2) 23,
3) (i) 95, (ii) 172, (iii) 5,
4) (i) 16, (ii) 34
5) (i) 0.1131 (ii) 0.3821

5.9 FITTING OF NORMAL DISTRIBUTION

Example: Fit a normal distribution of the data

X	F
0-10	2
10-20	4
20-30	10

30-40	3
40-50	1

Solution: Area Method:

we use area under standard normal curve to find expected frequencies. The steps of Area method are as under:

1. Find X and o of given data.
2. Write down the lower limit of each class and denote it by X.
3. For each value of X, find standard normal variate Z as:

   Z= $\frac{X-\bar{X}}{\sigma}$
4. Find the area under the normal curve from 0 to Z from table.
5. Find the area for each class interval. If Z's have same sign, then class areas are found by subtracting the successive areas and if Z's have opposite signs, then class areas are found by adding the successive areas.
6. Multiply the area calculated by N to obtain expected frequency of each class.

X	F	M.V	dx= X-25	d'x = dx/10	fd'x	fd'x2
0-10	2	5	-20	-2	-4	8
10-20	4	15	-10	-1	-4	4
20-30	10	25	0	0	0	0
30-40	3	35	10	1	3	3
40-50	1	45	20	2	2	4
	20				-3	19

$\overline{\times}= A + \frac{\sum fd\prime X}{\sum f} \times i$

$= 25 + \frac{3}{20} \times 10$

$= 25 - 1.5$

$= 23.5$

Standard Deviation ($\boldsymbol{\sigma}$) $= \sqrt{\frac{\sum fdx^2}{n} - \left(\frac{\sum fdx}{n}\right)^2} \times i$

$= \sqrt{\frac{19}{20} - \left(\frac{-3}{20}\right)^2} \times 10$

$= \sqrt{0.95 - 0.0225} \times 10$

$= \sqrt{0.9575} \times 10$

$= 0.9631 \times 10$

$= 9.63$

X	Lower limit	Z= $\frac{X-\bar{X}}{\sigma}$	Area from 0 to Z	Class probability	Exp. Freq = Probability * N
Below 0	-∞	-∞	0	0.5-0.4927= 0.0073	0.146 =0
0-10	0	-2.44	0.4927	0.4927- 0.4192= 0.0735	1.47 = 1
10-20	10	-1.40	0.4192	0.4192-0.1406= 0.2786	5.57 = 6
20-30	20	-0.36	0.1406	0.1406 + 0.2486= 0.3892	7.78 = 8
30-40	30	0.67	0.2486	0.2486- 0.4564 = 0.2078	4.16 = 4
40-50	40	1.71	0.4564	0.4564-0.4970= 0.0406	0.81 = 1
50 above	50	2.75	0.4970	0.5-0.4970 = 0.003	0.06 =0
					N = 20

5.10 KEYWORDS

- Normal Probability Curve: A normal curve is a bell-shaped curve, bilaterally symmetrical and continuous frequency distribution curve.
- Normal Probability Distribution: A continuous probability distribution for a variable is called as normal probability distribution or simply normal distribution. It is also known as Gaussian/ Gauss or LAPlace – Gauss distribution.
- Standard score: Standard score or z-score is a transformed score which shows the number of standard deviation units by which the value of observation (the raw score) is above or below the mean.

5.11 SUM UP

There are three types of probability distribution *i.e.* Binomial distribution, the Poisson distribution and normal distribution. The normal distribution, also called the Gaussian distribution, is a probability distribution commonly used to model phenomena such as physical characteristics (e.g. height, weight, etc.) and test scores. Due to its shape, it is often referred to as

the bell curve. When plotted on a graph, the data follows a bell shape, with most values clustering around a central region and tapering off as they go further away from the center. Normal distributions are also called Gaussian distributions or bell curves because of their shape. A normal curve can be used to determine the percent of the total area under the normal curve associated with any given sigma distance from the mean. Such a curve is formed as a result of plotting frequencies of scores of a continuous variable in a large sample. The curve is known as a normal probability distribution curve because its ordinates provide relative frequencies or probabilities instead of the observed frequencies. Standard score or z-score was also discussed in detail and the discussion covered concept, properties, uses and computation of z-score.

5.12 SUGGESTED READINGS

- Beri G.C. (2007), Business Statistics, (2nd ed.) New Delhi, Tata McGraw Hill.
- Levin, J. & Fox, J.A. (2006) Elementary Statistics in Social Research (10th ed.) India, Pearson Education.

DIPLOMA COURSE IN STATISTICAL ANALYSIS AND RESEARCH METHODOLOGY

SEMESTER II

SARM 4: TIME SERIES ANALYSIS AND PROBABILITY DISTRIBUTIONS

UNIT 6: INTERPOLATION AND EXTRAPOLATION

STRUCTURE

6.0 LEARNING OBJECTIVES

After Reading this Unit, Learner will be able to know:

- Meaning of interpolation and extrapolation
- Assumptions of validity
- About the methods of interpolation and extrapolation
- Advantages and Disadvantages

6.1 INTRODUCTION

In real-life economic situations, decision-making must be judicious, operative, and capable. It is well well-known fact that decision-making activity is extremely intricate given the kind and nature of economic variables. Decision-making surroundings, necessities information to scrutinize, comprehend, and develop an appropriate model for effective and efficient management. Information is critical to the decision-making process. Sometimes information relating to the decision ecosystem is available and sometimes essential information is either absent or not available for several reasons. Interpolation and Extrapolation are both statistical methods that enable the estimation of unfamiliar values in a given series. Both extrapolation and interpolation are valuable approaches to calculating or estimating the hypothetical values for an unidentified variable based on the reflection of other data points. Nevertheless, it can be difficult to distinguish among these techniques and comprehend how they differ from each other.

In the regular management of any industry or for doing forecasting for decision-making, information is gathered regularly. But it may be difficult or not possible to collect information for every time point. In the real-life world and business, several times we come across situations where we have to make an estimation of a value that is either missing or not available in the given series of information or forecast a prospective value. To handle these types of situations instead of being subject to some guesswork, the techniques of interpolation and extrapolation are quite helpful. For instance, the census of the populace in India takes place every 10 years, i.e., we have the survey statistics for 1951, 1961, 1971, 1981, 1991, and 2001. Taking the support of this accessible information, if anyone desires to know the survey data for the years 1996 or 2007 then with the help of the technique of interpolation and extrapolation the same can be estimated and arrived at. The requirement for interpolating missing information or making predictions or estimates arises

in some fields like economics, commerce, social sciences, actuarial work, population studies, etc. Therefore, the practice of interpolation and extrapolation are very supportive in assessing the missing values or forecasting future values. The present chapter is focused on Interpolation and Extrapolation.

6.2 MEANING AND DEFINITION OF INTERPOLATION AND EXTRAPOLATION

In simple words, interpolation is done to estimate a value inside the specified range of the series while extrapolation on the other hand, it deals with obtaining the prediction or estimates of the required information in the earlier or future outside the specified range of the series. Interpolation, therefore, talks about the insertion of an in-between value in a sequence of items while extrapolation denotes anticipating a value for the future. Every time the technique of interpolation or extrapolation is used, it is based on the supposition that the variable whose value is to be projected is the function of the other variable. A variable is said to be the function of the other, if for any values of the independent variable (say x) we can always find a certain value of the dependent variable (say y).

Explanation of the concept with an example: Let us assume that there are two variables x and y, x being the independent variable and y the dependent variable. Further, let the given values of x be X_0, X_1, X_2….. Xn, and let the resultant values of y be Y_0. Y_1, Y_2, …. Yn, respectively. If there is a requirement to guess the value of y, for any value of x between the limits, X_0 and this can be done by using the method of Interpolation. For instance, imagine we have been given the sale values figures for the years (x) 2010, 2012, 2013, 2015, and 2018 and we want to know the sale values for any year between 2010 and 2018, say, 2017, 2014, etc. This can be done by the technique of interpolation. On the other hand, if we have to estimate the sale values for the period outside the range 2010-2018, say, for 2008 or 2020, the method is known as extrapolation.

One of the easiest methods to distinguish these dissimilarities is to know the prefix of each term. Extra- denotes "in addition to," while inter- means "in between." Consequently, extrapolation points out a user is attempting to obtain a value in addition to available values, at the same time interpolation indicates that they want to ascertain a new value in between existing values.

“Interpolation is the estimation of a most likely estimate in given conditions. The Technique of estimating a past figure is termed as interpolation, while that of estimating a probable figure for

the future is called extrapolation." by M. Harper

6.3 ASSUMPTIONS OF INTERPOLATION AND EXTRAPOLATION

As has been stated above, the interpolation or extrapolation technique is applied based on particular suppositions. The following are some of the assumptions that are taken into consideration while using the techniques of interpolation and extrapolation.

i. **No sudden or violent fluctuations in the intervening period**: At the time of interpolating or extrapolating a value, it is at all times assumed that there are no sudden deviations in the given data. In other words, the values should communicate the periods of normal and steady economic state of affairs. To put it differently, the given data on which the interpolation or extrapolation technique is to be applied should be free from all types of anomalies and all categories of unsystematic and uneven variations. If, for instance, we are interpolating the data of sales figures of a company for the year 2020 and we are given the figures of sales data for the year 2017. 2018, 2019, and 2021 we would assume that the sales of the company under consideration have grown up evenly and there are no aggressive ups and downs in these sales figures. There are a number of cases like earthquakes, wars, floods, labor strikes, lockouts, economic boom depression and political disturbances, etc., which may lead to violent ups and downs in the values, which should not be considered while applying the techniques of interpolation or extrapolation.

ii. **The percentage of change of figures from one period to another is uniform**: The second supposition is that the degree of variation of the data is uniform. Therefore, in the example of sales data given above, if we want to interpolate or extrapolate the sales figure, it is assumed that the data from a period from 2017 to 2021 has witnessed evenly growth, i.e. free from all the types of abnormalities. Taking into account these assumptions, missing data can be interpolated with a reasonable degree of precision.

6.4 ACCURACY OF INTERPOLATION AND EXTRAPOLATION

As the interpolation and extrapolation techniques are based on particular postulations which may sometimes pose some difficulties in practice, the values so estimated, may not at all times be precise or dependable, and it is difficult to ascertain the degree of error of the estimate. So, the accuracy of the interpolated or extrapolated values is affected due to:

a) likely variations in the values of the trend under investigation, which is given by the existing information at our disposal.

b) A known fact around the sequence of happenings that may disturb the value of the observable fact under consideration. If it is known that the expected value of the specified event at a specific period is influenced by random circumstances, like political disturbances, floods, etc., then the interpolation or extrapolation is disturbed and these known facts should be taken into account while reaching a certain conclusion for making any estimation for missing values.

6.5 NEED AND IMPORTANCE OF INTERPOLATION AND EXTRAPOLATION

The methods of interpolation and extrapolation are of immense real-world use, because of:

(i) Non-availability of data: Interpolation may also be necessary in case the data are inadequate because of gaps in the data or are ineptly gathered while collecting the information.

(ii) Loss of data: Data from some of the periods may be deleted, damaged, or missing due to several causes like wrong management or random and natural causes like fire, floods, etc. Such types of data may be acquired with the help of the interpolation method. The interpolation technique is therefore supportive in filling up the data gaps in accessible data.

(iii) To estimate the intermediate values: Owing to several financial and organisational teething troubles, information may not be accumulated on a survey basis and random sampling practices may be used to find the appropriate data. The in-between differences are then satisfied by interpolation methods.

(iv) To bring uniformity in the data: From time to time, it so happens that the data relating to a particular event are assembled by diverse working groups working in different categories of groups and to draw any inferences from this data is difficult for evaluation. To achieve equality in the groups, the interpolation method is resorted to. If for instance, the information is gathered for two diverse dates, for doing a comparison in them, they have to be brought at one point in time. For instance, in a nation the population survey was done in 2020, and in India the survey was done in 2021. For doing a comparison between the populations of the two nations either India's population is to be interpolated for 2020 or the other nation's populace is to be projected by extrapolation for 2021.

(v) For doing forecasts: Projection of future data is a fundamental necessity in any policy formation or economic planning. The extrapolation technique is used in making predictions.

For instance, a company wants to project for the next financial year based on records. This can easily be done with the help of extrapolation technique.

(vi) To ascertain the positional averages in **continuous frequency spreading**: The interpolation technique has been used to develop the formulae for the working out of the median, quartiles, quintiles, cortiles, deciles, percentiles, and mode in case of continuous frequency distribution.

6.6 INTERPOLATION

It is a technique of fitting the data points to denote the value of a function. It has some applications in engineering, commerce, industry, and science that are used to build new data points within the range of a discrete data set of known data points or can be used for finalizing a formula of the function that will pass from the given set of points (x, y). In this study material, we will be discussing the concept of interpolation in Statistics, its formulas, and its uses in detail.

Interpolation is a technique of deriving a simple function from a particular discrete data set such that the function passes through the provided data points. This supports to conclusion of the data points in between the given data. This process is at all times required to figure out the value of a function for an in-between value of the independent function. To summarize, interpolation is a method to determine the unidentified values that lie in between the known data points. It is frequently used to forecast the unknown values for any ecologically connected data points such as noise level, rainfall, elevation, and so on.

Hirach "Interpolation is the art of understanding between the lines of the table."

Interpolation Formula

The unknown value on the data points can be found using the linear interpolation and Lagrange's interpolation formula.

The Linear interpolation formula is given by

$$y = y_1 + \frac{x - x_1}{x_2 - x_1} \times (y_2 - y_1)$$

Similarly, the Lagrange's interpolation formula is given as:

$$y = \frac{(x - x_1)(x - x_2).....(x - x_n)}{(x_0 - x_1)(x_0 - x_2).....(x_0 - x_n)} y_0 + \frac{(x - x_0)(x - x_2).....(x - x_n)}{(x_1 - x_0)(x_1 - x_2).....(x_1 - x_n)} y_1 + .. + \frac{(x - x_1)(x - x_1).....(x - x_{n-1})}{(x_n - x_0)(x_0 - x_1).....(x_n - x_{n-1})} y_n$$

Interpolation explained with an example

Interpolation means ascertaining a value from the existing values in a given data set. In other words, it is a process of putting in or interjecting a middle value between two other values.

In data science or mathematics, interpolation is something like calculating a function's value based on the value of other data points in a given sequence. This function may be represented as f(x), and the known x values may range from X_0 to X_n.

For instance, imagine we have a regression line y = 3x + 4. We know that, to produce this "best-fit" line, the value of x must be between 0 and 10. Supposing we choose x = 5 Based on this best-fit line and equation, we can estimate the value of y as the following:

$$y = 3(5) + 4 = 19$$

Our x value (5) is within the range of adequate x values used to make the line of finest fit, so this is a valid y value, which we have computed by interpolation.

6.7 INTERPOLATION METHODS

Three of the most common interpolation methods are the following:

- Linear Interpolation
- Polynomial Interpolation
- Spline Interpolation

1. Linear Interpolation

Linear interpolation is amongst the simplest interpolation techniques. At this point, a straight line is drawn amid two points on a graph to control the other unidentified values. This simple technique frequently produces wrong estimates.

2. Polynomial Interpolation

While using the polynomial interpolation method, polynomial roles are used on a graph to estimate the values in the set of data that has been misdirected. It is a somewhat more comprehensive, perfect method. The polynomial graph fills in the curve amongst identified points to find the missing data between those points.

There are multiple methods of polynomial interpolation:

- Lagrange interpolation
- Newton polynomial interpolation
- Spline interpolation

The Newton method is also identified as Newton's divided differences interpolation polynomial. The Lagrange and Newton interpolation techniques outcome in the smallest polynomial function, i.e., the polynomial of the lowermost potential point that goes across the data points in the data set. Both methods produce the same outcome but to arrive at the results both use different types of calculations.

3. Spline Interpolation: In spline interpolation, piecewise functions are employed to make an estimate of the missing values and fill the gaps in a data set. In its place of assessing one polynomial for the whole of the data set as takes place in the Lagrange and Newton methods, spline interpolation describes multiple simpler polynomials for subgroups of the data. For this purpose, it commonly delivers more precise results and is believed to be a more trustworthy method.

- **Nearest Neighbour Method**– This technique introduces the value of an interpolated point to the value of the most nearby data point. Consequently, this technique does not create any new data points.
- **Cubic Spline Interpolation Method**– This process fits a diverse cubic polynomial between each pair of data points for curves or between sets of three points for surfaces.
- **Shape-Preservation Method**– This method is also known as Piecewise Cubic Hermite Interpolation (PCHIP). It maintains the monotonicity and the shape of the data. It is for curves only.
- **Thin-plate Spline Method**– This technique contains smooth surfaces that also extrapolate well. It is only for surfaces only
- **Biharmonic Interpolation Method–** This method is applied to the surfaces only.

6.8 EXTRAPOLATION

In Statistics, **Extrapolation** is a method of assessing the value outside the different range of the specified variable based on its connection with another variable. It is a very essential notion not only in Mathematics but also in other fields like Psychology, Sociology, Statistics, etc., with some definite data. Now, we will examine in detail regarding definition, formula, and examples of

extrapolation. Another more significant concept is an **interpolation,** which has been discussed above as it is an estimation between the given data.

Extrapolation is described as an estimation of a value based on expanding the identified series or factors outside the range that is known. In other words, extrapolation is a method in which the data values are studied as points such as x_1, x_2, …., x_n. It normally occurs in statistical data very regularly, if that data is sampled intermittently and it approximates the next data point. One such instance is when a driver is driving a car, he ordinarily **extrapolates** about road conditions beyond his vision.

Extrapolation is a statistical method that is used in comprehending the unidentified data from the known data. It tries to forecast future data based on past data. For instance, estimating the size of the population of a country for policy making by the government after a few years based on the existing population size and its rate of growth. Another example is forecasting the sale of a particular product in the future based on the past sales record of a company.

6.9 EXTRAPOLATION METHODS

Extrapolation is categorized into three types, namely

- Linear extrapolation
- Conic extrapolation
- Polynomial Extrapolation

Let us briefly talk about these three kinds of extrapolation methods.

1. Linear Extrapolation

For any linear function, the linear extrapolation method delivers a good result when the point to be projected is not excessively far off from the given data. It is typically done by sketching the tangent line at the endpoint of the given graph and that will be extended beyond the limit.

2. Conic Extrapolation

A conic section can be formed with the assistance of five points closer to the end of the given i.e. known data. The conic segment will curve back on itself if it is a circle or ellipse. But for parabola or hyperbola, the curve will not back on itself as it is relative to the X-axis.

3. Polynomial Extrapolation

A polynomial curve can be shaped with the assistance of the whole of the identified data or near the endpoints. This technique is normally performed using Lagrange interpolation or Newton's system of finite series that arranges for the data. The final polynomial is used to extrapolate the data using the connected endpoints.

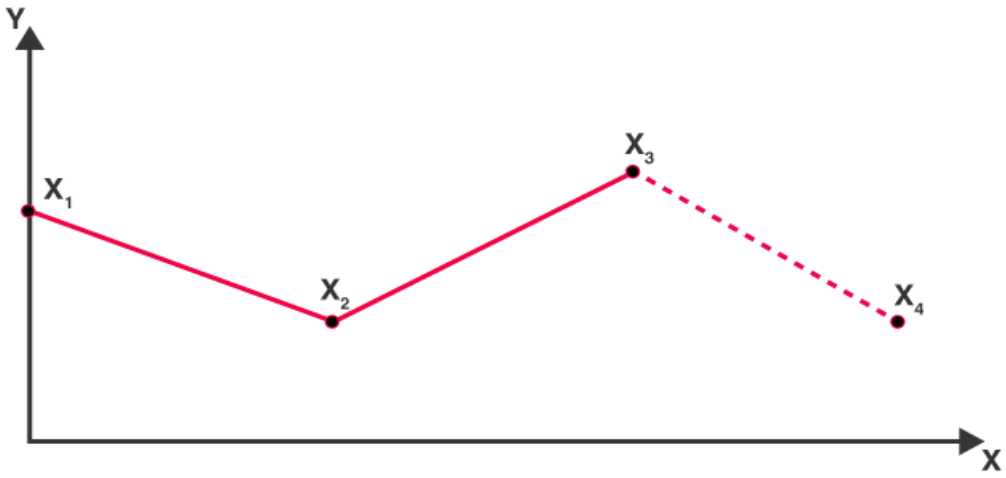
Extrapolation Formula

Let us consider the two endpoints in a linear graph (x_1, y_1) and (x_2, y_2) where the value of the point "x" is to be extrapolated, and then the extrapolation formula is given

$$y(x) = y_1 + \frac{x - x_1}{x_2 - x_1}(y_2 - y_1)$$

Extrapolation Graph

As it is known, extrapolation is a method of forecasting the data point about the exterior of a curve when normally a few points are given. In the model given below, the identified data are x_1, x_2, x_3. Locating the point x_4 is acknowledged as an extrapolation point.



6.10 ADVANTAGES AND DISADVANTAGES OF INTERPOLATION

Advantages of Interpolation

- Can assess extreme changes in terrain such as Cliffs and Fault Lines.
- Thick evenly spaced points are well interpolated (flat areas with cliffs).

- Can augment or decline the amount of sample points to influence cell values.

Disadvantages of Interpolation

- Cannot make an estimation above maximum or below minimum values.
- Not very good for peaks or mountainous areas.

6.11 ADVANTAGES AND DISADVANTAGES OF EXTRAPOLATION

Advantages of Extrapolation

- Extrapolation is the analysis of data based on past trends. As past data is easily available based on which forecasting can be done to make informed decisions.
- It is an uncomplicated technique of forecasting as the rough judgment of the data can be done based on the earlier data.
- Not much data information is required as past data of a very near period is relevant for extrapolating the data.
- It is prompt and low-priced as the cost involved in collecting the past data is not high.
- It can encourage staff if aims are high. With the help of extrapolation, the staff can be encouraged by educating them that the targets have been fixed on a scientific basis.

Disadvantages of Extrapolation

- It can be uneven if there have been no changes in rise and fall with past data
- It undertakes that historical trends will always carry on into the future, which is doubtful in numerous business environments. The results will not be accurate, if there is a change in the policy of the government about a particular industry. There might be chances that the circumstances may change on a global basis like war or abnormal currency fluctuations etc.
- The technique of extrapolation overlooks the qualitative factors which cannot easily be quantifiable. If for instance, there is a change of fashion or liking of the consumers in the clothing industry, it will be difficult to extrapolate the data based on records. Extrapolation done based on the past data may not produce accurate results.
- It is a possibility that high-pitch targets can strain staff members. If targets are too high or too low based on the forecast of the previous data, it can promote dissatisfaction among the labour. As high targets will unnecessarily pressure them and low targets will result in suboptimal use of the resources.

- Extrapolating beyond a reasonable range is quite a difficult task. Sometimes a company wants to modernize the machinery to increase its productivity. It may be difficult to do a cost-benefit analysis based on future production.

6.12 APPLICATION OF EXTRAPOLATION AND INTERPOLATION

Interpolation time and again delivers a legitimate estimation of an unknown value, because of this, it's deliberated as a more dependable assessment method than extrapolation. Both approaches are advantageous for different purposes. Interpolation is particularly valuable to guess missing values or lost records to complete the records for deciding on doing any project or activity. Extrapolation is done to make forecasts about an event or occurrence based on a set of known or past values. In the real world, interpolation and extrapolation are applied in numerous fields, including the following:

- **Mathematics** to ascertain function values to reveal unidentified values to solve real-world problems;
- **Science** to make weather prediction models, forecast rainfall, or predict unknown chemical concentration values; and
- **Statistics** to forecast prospective data, such as population growth or the spread of a disease.

6.13 COMPARISON BETWEEN INTERPOLATION AND EXTRAPOLATION

Interpolation	Extrapolation
The interpretation of the values between two points in a data set. It is the prediction of a most likely assessment in the given circumstances.	Assessing a value that's outside the data set. Assessing a likely figure for the future is called extrapolation.
Predominantly used to ascertain missing past values. When data from some past periods are missing, information connecting to such projects may be assessed to finish the records by interpolation	Performs a most important role in predicting. It plays an important part in economic forecasting. For financial forecasting, prediction of future data is indispensable. This is done by extrapolation.

The expected information is more likely to be accurate. Interpolation has a preference because it has a better probability of finding an acceptable assessment.	The projected values are only possibilities, so they may not be completely accurate. In extrapolation, we normally assume that our perceived trend lasts for values of x outside the range. We worked to form our model. This may not be the case. So proper care should be taken while doing extrapolation.
It can be computed graphically. It is one of the easiest methods of interpolation.	The graphic method is not useful for extrapolation.
The technique of estimating a past figure is termed interpolation.	The technique of estimating a Future figure is termed as interpolation.

6.14 SUM UP

It is well known fact that decision-making activity is extremely intricate given the kind and nature of economic variables. Information is critical to the decision-making process. The number of times information relating to the decision ecosystem is available and sometimes essential information is either absent or not available for several reasons. Interpolation and Extrapolation are both statistical methods that enable the estimation of unfamiliar values in a given series. In simple words, interpolation is done to estimate a value inside the specified range of the series while extrapolation on the other hand deals with obtaining the prediction or estimates of the required information in the earlier or future outside the specified range of the series. At the time of interpolating or extrapolating a value, it is at all times assumed that there are no sudden deviations in the given data. The second supposition is that the degree of variation of the data is uniform. Interpolation and extrapolation are useful in the non-availability of data and loss of data, for estimating the intermediate values, bringing uniformity in the data, doing forecasts and ascertaining the positional averages in continuous frequency spreading.

6.15 QUESTIONS FOR PRACTICE

A. Short Answer Type Questions

Q1.Write a short note on interpolation with an example.

Q2.Write a short note on extrapolation with an example.

Q3.What are the different methods of interpolation?

Q4.What are the main differences between interpolation and extrapolation?

B. Long Answer Type Questions

Q1.What are the applications of interpolation and extrapolation in forecasting information for businesses?

Q2.What is the Need and Importance of Interpolation and Extrapolation?

Q3.What is extrapolation? Discuss its various methods.

Q4.What are the advantages and disadvantages of interpolation and extrapolation?

Q5.Why Interpolation and Extrapolation are required to be done? Discuss its importance.

Q6.Discuss the important assumption of interpolation and extrapolation

6.16 PRACTICAL QUESTIONS WITH SOLUTION

Q1. If (20, 60) and (40, 100) are the two points on a straight line, find the value of y, when x = 60 using linear extrapolation.

Solution: Given x1=20 and y1=60 similarly x2= 40 and y2=100, x=60

We know that $y(x) = y_1 + ((x - x_1) / (x_2 - x_1) (y_2 - y_1))$.

Now putting the value from the above we get

$$y(60) = 60 + \frac{(60 - 20)}{(40 - 20)} \times (100 - 60)$$

On solving the equation, we get $y(60) = 60 + 2 \times 40$

Thus, by solving we get $y(60) = 140$

Q2. If (60, 30) and (80, 90) are two points on a straight line, find the value of 'y' when x = 120 using linear extrapolation.

Solution: Given x1=60 and y1=30 similarly x2= 80 and y2=90, x=120

We know that $y(x) = y_1 + ((x - x_1) / (x_2 - x_1) (y_2 - y_1))$.

Now putting the value from the above we get

$$y(120) = 30 + \frac{(120-60)}{(80-60)} \times (90-30)$$

On solving the equation, we get $y(120) = 30 + 3 \times 60$

Thus, by solving we get y (120) = 210.

Q3. Find the value of y when x=10 by Lagrange's interpolation method.

x	**5**	**6**	**9**	**11**
y= f(x)	**12**	**13**	**14**	**16**

Solution:

Firstly, we will write Lagrange's interpolation method formula as given below.

$$\frac{(x-x_1)(x-x_2)(x-x_3)}{(x_0-x_1)(x_0-x_2)(x_0-x_3)} \times f(x_0) + \frac{(x-x_0)(x-x_2)(x-x_3)}{(x_1-x_0)(x_1-x_2)(x_1-x_3)} \times f(x_1) + \frac{(x-x_0)(x-x_1)(x-x_3)}{(x_2-x_0)(x_2-x_1)(x_2-x_3)} \times f(x_2) +$$

$$\frac{(x-x_0)(x-x_1)(x-x_2)}{(x_3-x_0)(x_3-x_1)(x_3-x_2)} \times f(x_3)$$

The given value of x and y are depicted in the table

x	$x_{0=}$ 5	x_1= 6	x_2= 9	x_3= 11
y= f(x)	12	13	14	16

Putting the values in the formula we get f(x)=

$$\frac{(x-6)(x-9)(x-11)}{(5-6)(5-9)(5-11)} \times 12 + \frac{(x-5)(x-9)(x-11)}{(6-5)\,6-9)(6-11)} \times 13 + \frac{(x-5)(x-6)(x-11)}{(9-5)(9-6)(9-11)} \times 14 +$$
$$\frac{(x-5)(x-6)(x-9)}{(11-5)(11-6)(11-9)} \times 16$$

Now f (10) $= \frac{(10-6)(10-9)(10-11)}{(5-6)(5-9)(5-11)} \times 12 + \frac{(10-5)(10-9)(10-11)}{(6-5)(6-9)(6-11)} \times 13 + \frac{(10-5)(10-6)(10-11)}{(9-5)(9-6)(9-11)} \times 14$
$+ \frac{(10-5)(10-6)(10-9)}{(11-5)(11-6)(11-9)} \times 16$

Solving the equation, we get

$$\frac{(4)(1)(-1))}{(-1)(-4)(--6)} \times 12 + \frac{(5)(1)(-1)}{(1)(-3)(-5)} \times 13 + \frac{(5)(4)(-1)}{(4)(3)(-2)} \times 14 + \frac{(5)(4)(1)}{(6)(5)(2)} \times 16$$

=14.666.

Hence the value of y=f(x) will be 14.666 when x=10.

- Interpolate the value f(x) when x=4 in the following table.

x	3	5	7	9
y =f(x)	180	150	120	90

Solution:

x	$x_0=3$	$x_1=5$	$x_2=7$	$x_3=9$
y =f(x)	180	150	120	90

As the value of x lies between x_0 and x_1

First, we calculate the value of 'h', which is h=2.

Then $u= \frac{x-x_0}{h}$, where x=4 and x_0 = 3 hence $u= \frac{4-3}{2} = \frac{1}{2} =.05$.

Difference Table

x	y	Δy	Δ^2 y	Δ^3 y
3	180			
		-30		
5	150		0	
		-30		0
7	120		0	
		-30		
9	90			

The formula for interpolating the above problem is

$$y=y_0 + \frac{u}{1!}\Delta y_0 + \frac{u(u-1)}{2!}\Delta^2 y_0 + \frac{u(u-1)(u-2)}{3!}\Delta^3 y_0$$

$$y = 180 + \frac{(.05)}{1!}(-30) + \frac{(.05)(.05-1)}{2!}(0) + 0$$

$$y = 180 - 15 = 165$$

Hence y=165 is the answer.

Q4. Estimate the turnover of business enterprises for the year 2021 from the under mentioned data.

Year	**2016**	**2017**	**2018**	**2019**	**2020**	**2021**
Turnover in Rs. lacs	**50**	**?**	**100**	**150**	**260**	**?**

Solution:

Year	Turnover In Rs. lacs	
2016	50	y_0

2017	?	y_1
2018	100	y_2
2019	150	y_3
2020	260	y_4
2021	?	y_5

As the known values are 4

$\therefore (y - 1)^4 = 0$

$y_4 - 4y_3 + 6y_2 - 4y_1 + y_0 = 0$ …… (1)

The second equation can be calculated by increasing the suffixes of each term of 'y' by one and let the coefficients same.

then we get

$y_5 - 4y_4 + 6y_3 - 4y_2 + y_1 = 0$ ….(2)

Substitute y values in equation (1) we get

$y_4 - 4y_3 + 6y_2 - 4y_1 + y_0 = 0$

$260 - 4(150) + 6(100) - 4y_1 + 50 = 0$

$260 - 600 + 600 - 4y_1 + 50 = 0$

$-4y_1 + 310 = 0$

$y_1 = \frac{-310}{-4} = 77.50$

Hence sales in 2017 are Rs.77.50 lacs.

Now substitute y values in equation (2) we get

$y_5 - 4y_4 + 6y_3 - 4y_2 + y_1 = 0$

$y_5 - 4 (260) + 6 (150) - 4 (100) + 77.50 = 0$

$y_5 - 1040 + 900 - 400 + 77.50 = 0$

$y_5 - 462.50 = 0$

$y_5 = 462.50$

Hence sales in 2021 are Rs.462.50 lacs

6.17 MCQs

1. Interpolation of the values is done by

a) Correlation analysis
b) Regression Analysis
c) Both a and b above
d) Curve fitting and regression analysis

2. Linear Interpolation is:

a) Easy and accurate
b) Easy
c) Precise
d) None of the above

3. Which of the following is more costly to implement?

a) Linear interpolation
b) Polynomial & linear interpolation
c) Polynomial interpolation
d) None of the above

4. Interpolation is a system used for:

a) Discovering the lost values
b) Finding the link between two variables
c) Matching the two series
d) Finding the most probable missing relations

5. From the under-mentioned, which is the best common formula for interpolation and extrapolation is given by:

a) Lagrange
b) Newton and Gauss
c) Bernoulli
d) Newton and Gregory

6. Drawing a curvature line at the ending of the known figures and expanding it outside that boundary is named as

a) Dentipolation
b) Extrapolation
c) Antipolation
d) Interpolation

7. Interpolation is not affected by

a) Uneven Variations
b) Unanticipated Events
c) Violent Fluctuations
d) None of the above

8. Graphical technique can be used for

a) Interpolation and extrapolation
b) Interpolation only
c) Extrapolation only
d) None of the above

9. If there are sequential missing values in a series, their educated guess is

a) A difficult problem
b) Cannot be done
c) Result cannot be relied upon
d) All of the above

10. Which of the under mentioned is the easiest technique of Estimating?

a) Extrapolation
b) Regression
c) Exponential smoothing
d) Moving average method

Answers of MCQ

1	2	3	4	5	6	7	8	9	10
d	b	c	d	a	b	d	a	b	a