Course Code:- MGO-6104

Course Name:- Business Statistics

MASTER OF BUSINESS ADMINISTRATION (Management Studies)

PROGRAMME DESIGN COMMITTEE

Prof. Masood Parveez
Vice Chancellor – Chairman
MTSOU, Tripura

Prof. Abdul Wadood Siddiqui
Dean Academics
MTSOU, Tripura

Prof. C.R.K. Murty
Professor of Distance Education
IGNOU, New Delhi

Prof. Mohd. Nafees Ahmad Ansari
Director of Distance Education
Aligarh Muslim University, Aligarh

Prof. Ashish Mishra
Professor of Management
Mangalayatan University, Jabalpur

Prof. Rajeev Sharma
Professor of Management
Mangalayatan University, Aligarh

Prof. K. Ravi Sankar
Professor of Management
IGNOU, New Delhi

Prof. Anurag Saxena
Professor of Management
IGNOU, New Delhi

Prof. Arvind Hans
Professor of Management
MTSOU, Tripura

Prof. Prakash D Achari
Professor of Management
Usha Martin University, Ranchi

COURSE WRITERS

Dr Arvind Hans
Professor of Management
MTSOU, Tripura
MGO-6101 Principles and Practices of Management

Dr Saifullah Khalid
Associate Professor of Management
MTSOU, Tripura
MGO-6102 Managerial Economics

Dr Meenakshi Kumari
Assistant Professor of Management
MTSOU, Tripura
MGO-6103 Accounting for Managers

Dr. Nyarik Geyi
Assistant Professor of Management
MTSOU, Tripura
MGO-6104 Business Statistics

Mr. Rana Taku
Assistant Professor of Management
MTSOU, Tripura
MGO-6105 Communication for Management

COURSE EDITORS

Prof. Ashish Mishra
Professor of Management
Mangalayatan University, Jabalpur

Prof. Rajeev Sharma
Professor of Management
Mangalayatan University, Aligarh

Prof. Anurag Saxena
Professor of Management
IGNOU, New Delhi

Dr. Swati Saxena
Associate Professor of Management
Mangalayatan University, Jabalpur

Dr. Praksh Mishra
Associate Professor of Management
Mangalayatan University, Jabalpur

Dr. Ashutosh Saxena
Associate Professor of Management
Mangalayatan University, Jabalpur

FORMAT EDITORS

Dr. Nitendra Singh
Associate Professor of English
MTSOU, Tripura

Ms. Angela Fatima Mirza
Assistant Professor of English
MTSOU, Tripura

Dr. Faizan
Assistant Professor of English
MTSOU, Tripura

Ms. Vanshika Singh
Assistant Professor of English
MTSOU, Tripura

MATERIAL PRODUCTION

1. Mr. Himanshu Saxena
2. Ms. Rainu Verma
3. Mr. Jeetendra Kumar
4. Mr. Khiresh Sharma
5. Mr. Ankur Kumar Sharma
6. Mr. Pankaj Kumar

CONTENT

Page No.

Block I: Introduction **5-112**

Unit 1: Meaning and definitions of Statistical Data; Applications of Statistics in Managerial decision making;
Unit 2: Frequency Distributions; Measures of Central Tendency: Mean, Median, Mode and their implications;
Unit 3: Measures of Dispersion: Range, Quartile Deviation, Mean Deviation and Standard Deviation.

Block II: Correlation and Regression **113-182**

Unit 4: Meaning and uses of correlation
Unit 5: Meaning and uses of regression.
Unit 6: Various methods of calculation of the Coefficient of correlation and their analysis (Two Variable)
Unit 7: Regression analysis.

Block III: Analysis of Time Series **183-254**

Unit 8: Concept; Additive model, Multiplication model,
Unit 9: Seasonal variation, Cyclical Variation;
Unit 10: Various methods of Time Series Analysis and their applications in business.

Block IV: Probability **255-312**

Unit 11: Concept, its uses in business decision-making,
Unit 12: Addition and multiplication theorem of probability; Binomial theorem and its applications
Unit 13: Probability Distribution: Concept, applications of Binomial, Poisson and Normal Distributions.

Block V: Estimation Theory and Hypothesis Testing **313-409**

Unit 14: Estimation Theory and Hypothesis Testing: Sampling theory; Formulation of Hypotheses;
Unit 15: Application of Z-test, t-test,
Unit 16:F-test and ANOVA
Unit 17:Chi-Square test. Techniques of association of Attributes & Testing.
Unit 18: Chi-Square Test (□2)

BLOCK I: INTRODUCTION

UNIT 1: MEANING AND DEFINITIONS OF STATISTICAL DATA; APPLICATIONS OF STATISTICS IN MANAGERIAL DECISION-MAKING

Structure

1.0 Introduction of statistics

1.1 Objectives

1.2 Types of Statistics

1.3 Characteristics of Statistics

1.4 Scope of Statistics

1.5 Applications of Statistics in Managerial decision making

1.6 Limitations of Statistics

1.7 Let Us Sum Up

1.8 Key Words

1.9 Answers to Check Your Progress

1.10 Terminal Questions

1.0 INTRODUCTION OF STATISTICS

The interpretation of the term "statistics" might vary among individuals. The application of knowledge in this area is used in many ways within the context of everyday human existence. Statistics has been used for both personal and professional purposes. In everyday life, statistics are used for general computations.

Regarding the topic of home budgeting, In the realm of information, two distinct categories may be identified: quantitative and qualitative.

The data provided is of a qualitative nature. Therefore, individuals use this field of study to make informed decisions.

This analysis focuses on the issues pertaining to budgetary concerns, taking into account both forms of information available.

and analysis of data. It involves the use of various techniques and methods to summarize and interpret data, making it a valuable tool in many fields such as economics, psychology, and biology. Statistics allows researchers to draw conclusions and make informed decisions based on the information gathered from a sample of data. The field of statistics encompasses several key concepts and principles. One important aspect is the collection of data, which involves gathering information from a specific population or sample. This data is then organized and classified into different categories or variables, allowing for a systematic analysis. Another crucial component of statistics is

The process of tabulating, analyzing, interpreting, and presenting data. Several prominent scholars

The notion of Statistics being regarded as an independent mathematical discipline, apart from its classification as a branch of mathematics, is worth considering. In the present discourse, the user has expressed their thoughts and opinions on a particular subject matter

On the contrary, several scientific inquiries rely on data, hence emphasizing the significance of statistics in these investigations.

The use of data in the context of uncertainty and decision-making in the presence of uncertainty.

In contemporary society, characterized by the prevalence of computers and information technology, the significance of statistics is widely acknowledged across several academic areas. The field of statistics has its roots in the study of states and has gradually expanded its applications to include domains like as agriculture, economics, and commerce.

The fields included include biology, medicine, industry, planning, and education, among others. As at the present moment, there exists no other individual of the human species.

The walk of life is a domain in which the use of statistics is not feasible.

Meaning and definition

The terms 'Statistics' and 'Statistical' both originate from the Latin word 'Status', which pertains to a political state. The emergence of statistics as a separate field within the scientific method has occurred very recently.

The field of statistics encompasses the use of scientific methodologies for the purpose of gathering, arranging, condensing, displaying, and examining data, with the ultimate goal of drawing accurate inferences and reaching sound judgments based on this analysis. Statistics is a field of study that focuses on the methodical gathering of data.

The analysis and comprehension of quantitative information. In academic discourse, the term "statistic" is used to denote

1. Numerical data, such as population figures for certain geographic regions.
2. The field of inquiry concerned with the methodologies used to gather, analyze, and interpret empirical data.

The definition of statistics varies across writers and has evolved throughout time. In previous eras, the field of statistics was mostly limited to matters pertaining to the state. However, in contemporary times, it has expanded its scope to include almost all domains of human activity. Consequently, certain antiquated definitions that were limited in scope

The sphere of inquiry has been substituted with other definitions that provide a greater level of comprehensiveness and inclusivity.

The analysis conducted was comprehensive and thorough. Secondly, statistics may be defined in two distinct ways, namely statistical data and statistical methods. The below statements include many definitions of statistics in the context of numerical data.

Statistics are a collection of categorized data that depict the situation of individuals within a certain population. Specifically, they are the empirical data that may be expressed quantitatively via numerical values or organized in the form of tables or any other systematic organization.

Statistics refer to the quantification, enumeration, or estimation of natural phenomena, often via the use of numerical measures.

The data is organized in a methodical manner, analyzed thoroughly, and presented in a way that highlights significant interconnections.

According to A.L. Bowley: Statistics are numerical statement of facts in any department of enquiry placed in relation to each other.

Statistics may be called the science of counting in one of the departments due to Bowley, obviously this is an incomplete definition as it considers only the aspect of collection and ignores other aspects such as analysis, presentation and interpretation.

Bowley gives another definition for statistics, which states ‘statistics may be rightly called the scheme of averages’. This definition is also incomplete, as averages play an important role in understanding and comparing data and statistics provide more measures.

Definition by Croxton and Cowden: Statistics may be defined as the science of collection, presentation analysis and interpretation of numerical data from the logical analysis. It is clear that the definition of statistics by Croxton and Cowden is the most scientific and realistic one. According to this definition there are four stages:

➢ Collection of Data: It is the first step and this is the foundation upon which the entire data set. Careful planning is essential before collecting the data. There are different methods of collection of data such as census,

Sampling, primary, secondary, etc., and the investigator should make use of correct method.

➢ Presentation of data: The mass data collected should be presented in a suitable, concise form for further analysis. The collected data may be presented in the form of tabular or diagrammatic or graphic form.

➢ Analysis of data: The data presented should be carefully analysed for making inference from the presented data such as measures of central tendencies, dispersion, correlation, and regression etc.,

➢ Interpretation of data: The final step is drawing conclusion from the data collected. A valid conclusion must be drawn on the basis of analysis. A high degree of skill and experience is necessary for the interpretation.

Definition by Horace Secrist: Statistics may be defined as the aggregate of facts affected to a marked extent by multiplicity of causes, numerically expressed, enumerated or estimated according to a reasonable standard of accuracy, collected in a systematic manner, for a predetermined purpose and placed in relation to each other.

1.1 OBJECTIVES

After studying this unit, you should be able to:

- Understand meaning and characteristics of statistics
- Identify the scope of statistics
- Describe Applications of Statistics in Managerial decision making
- Limitations of Statistics

1.2 TYPES OF STATISTICS

Statistics may be classified into two primary categories: descriptive statistics and inferential statistics. These two classifications fulfill distinct objectives and are used across diverse disciplines for the examination and comprehension of data. Herein is a comprehensive exposition of the many categories:

1. Descriptive statistics.

Descriptive statistics serve the purpose of summarizing and providing a comprehensive description of the primary characteristics inherent in a given dataset. Data visualization tools

aid in the process of simplifying extensive quantities of data, hence enhancing its comprehensibility. Descriptive statistics encompasses a range of commonly used approaches and measurements.

Measures of central tendency include statistical measures that provide a description of the central location or average of a given dataset. The three most often used measures in statistical analysis are the mean (also known as the average), the median (representing the middle value), and the mode (indicating the value that occurs most frequently).

Measures of dispersion are statistical metrics that assess the extent of spread or variability within a dataset. Frequently used metrics include range, variance, standard deviation, and interquartile range.

Frequency distributions are visual representations, such as tables or graphs, that depict the occurrence of various values or categories within a dataset. Commonly used graphical tools for displaying frequency distributions include histograms, bar charts, and pie charts.

Measures of position include percentiles and quartiles, which provide insights into the comparative placement of a data item within a given dataset.

2.Inferential statistics

The field of inferential statistics encompasses the methods and techniques used to draw conclusions and make inferences about a population based on a sample Inferential statistics are used to derive conclusions and make inferences on a population by using a representative subset of data, known as a sample. These methodologies include the use of probability and sampling methods to formulate predictions and assess hypotheses. Inferential statistics encompasses a range of commonly used methods and ideas.

Hypothesis testing encompasses the process of constructing null and alternative hypotheses, followed by the use of statistical tests to ascertain the presence of a notable disparity or impact within the dataset.

Confidence intervals are statistical tools that provide a range of values in which a population parameter is expected to lie, accompanied with a predetermined degree of confidence.

Regression analysis is a statistical technique used to examine the link between variables. For instance, linear regression is utilized to investigate the linear relationship between two variables.

The statistical method known as Analysis of Variance (ANOVA) is used to assess and compare means across several groups, with the aim of identifying any statistically significant differences that may exist.

Probability Distributions: Different probability distributions, including the normal distribution, binomial distribution, and Poisson distribution, are used in order to represent and comprehend data.

Furthermore, apart from the aforementioned primary classifications, statistics may be further subdivided into several specific fields and methodologies, including but not limited to multivariate statistics, non-parametric statistics, Bayesian statistics, time series analysis, and other related areas. The selection of a statistical methodology is contingent upon the particular research or analysis goals and the characteristics of the data under investigation.

1.3 CHARACTERISTICS OF STATISTICS

Statistics is a discipline within the realm of mathematics and an academic domain that encompasses the systematic gathering, examination, interpretation, demonstration, and arrangement of

data. It assumes a pivotal function across a multitude of disciplines, including scientific, economic, social science, and commercial domains. The attributes of statistics encompass:

The process of data collecting in statistics involves the acquisition of data by several methodologies, including surveys, experiments, observations, and other applicable techniques.

Numerical Representation: In the field of statistics, data is often provided in numerical format, allowing for mathematical analysis and manipulation.

Empirical: Statistical data is derived from observations and measurements obtained from direct experience and real-world observations, as opposed to being derived from theoretical assumptions.

Universality: The use of statistics spans across several disciplines and domains, rendering it a versatile instrument for the purposes of decision-making and analysis.

Variability is a common characteristic seen in statistical data, characterized by the presence of randomness. The field of statistics provides tools and techniques to quantify and comprehend this variability. Measures such as variance and standard deviation are used to assess and gain insights into the extent of variability present in the data.

Objectivity: The field of statistics places a strong emphasis on the principles of objectivity and impartiality when it comes to the collecting and analysis of data. This is done in order to mitigate any potential biases and subjective judgments that may arise.

The process of aggregation in statistics often entails combining different data points to get summary statistics, such as means, medians, and percentages.

Interpretation: The process of interpreting statistical findings involves deriving conclusions, making predictions, or informing choices based on the obtained data. The comprehension of probability and uncertainty is necessary for this interpretation.

Comparison: Statistics enables the examination of diverse data sets, facilitating the evaluation of associations, disparities, and patterns.

Generalization: The field of statistics facilitates the process of generalizing research results obtained from a sample to a broader population, on the assumption that the sample is representative of such population.

Estimation: The field of statistics encompasses several techniques that enable the estimation of population parameters based on sample data. These techniques include the calculation of point estimates and the construction of confidence intervals.

Hypothesis testing is a statistical procedure used to evaluate hypotheses on the relationships between variables, with the aim of determining if there exists a statistically significant link or effect.

Data visualization is a commonly used technique in the field of statistics, whereby graphical elements such as graphs, charts, and other visual representations are employed to effectively communicate information.

Probability theory plays a crucial role in the field of statistics, serving as a foundational framework for many statistical approaches and ideas, particularly in addressing situations involving uncertainty.

The field of statistics places great emphasis on achieving accuracy and precision in the analysis and reporting of data. This commitment ensures that the obtained findings are both dependable and relevant.

In the field of statistics, it is important to acknowledge that the discipline is not characterized by exactness, but rather by a certain level of uncertainty and potential for mistake. Although statistics may provide vital insights, it is crucial to recognize its inherent limitations in terms of precision and accuracy.

The field of statistics undergoes continuous evolution as it responds to advancements in technology and methodology, enabling it to effectively incorporate new data sources and analytical approaches.

The combination of these attributes makes statistics a potent instrument for comprehending and formulating judgments grounded on facts, regardless of the domain, whether it scientific inquiry, commercial enterprises, governmental affairs, or other diverse sectors.

1.4 SCOPE OF STATISTICS

Statistics is a discipline within the field of mathematics that plays a crucial role in the process of gathering, analyzing, interpreting, and communicating data. The scope and use of this concept are extensive and include several disciplines, including as the natural sciences, commerce, social sciences, and governance. The following are essential elements pertaining to the extent and use of statistics:

The process of data collecting in the field of statistics encompasses the methodical gathering of information by various means such as surveys, experiments, observations, and other applicable techniques. The data may be categorized into two types: quantitative data, which consists of numerical values, and qualitative data, which comprises descriptive information. This data is fundamental for doing statistical analysis.

Data analysis is a crucial step in the research process, as it allows for the summarization, organization, and interpretation of acquired

data via the use of statistical methods. Descriptive statistics, including measures such as the mean, median, and standard deviation, provide a concise summary of the key attributes shown by a dataset.

Inferential statistics pertains to the process of drawing conclusions or making predictions about a population by using a sample of data. Methods like as hypothesis testing and confidence intervals are used to derive inferences about a population from a restricted sample.

Probability theory is a crucial component of statistical analysis, serving as the mathematical framework for comprehending and quantifying uncertainty and randomness. It assumes a pivotal role in domains such as risk assessment, forecasting, and decision-making.

Statistical modelling is a widely used approach utilized by statisticians to describe data and establish correlations between variables. Linear regression, logistic regression, and time series models are often used methodologies for the purpose of modelling real-world events.

Quality control is a crucial aspect in both manufacturing and service sectors, where statistical methods are used to ensure that goods or processes adhere to predetermined standards and regulations.

Research and scientific studies rely on the use of statistics for several purposes, including experimental design, data analysis, and the formulation of findings. This reliance is seen across diverse disciplines like as medicine, biology, psychology, and economics.

In the field of economics, economists apply statistical methods to examine and analyze various economic phenomena, including but not limited to economic trends, inflation, unemployment rates, and consumer behavior. This information has significant importance in the formulation of government policies and the making of commercial decisions.

The use of statistics in the field of business and marketing encompasses several applications such as market research, customer segmentation, demand forecasting, and performance assessment. A/B testing is a widely used technique for enhancing marketing strategy.

In the field of social sciences, researchers such as sociologists, psychologists, and political scientists use statistical methods to examine and derive insights from surveys and observational data pertaining to human behavior and social phenomena.

The use of statistics by governments encompasses several domains such as census data collection, public health surveillance, crime analysis, and resource allocation. The information provided plays a crucial role in shaping policy choices and determining the distribution of resources.

The use of statistical methods is prevalent in the field of finance and investment, where it serves as a crucial tool for evaluating risk, managing portfolios, and conducting research of the stock market.

Sports analytics is a field that recognizes the substantial impact of statistics on many aspects of sports, including the study of individual performance as well as the formulation of team strategies.

In the field of Environmental Science, statistical analysis is used by scientists to examine and interpret data pertaining to climate change, pollution, and the well-being of ecosystems.

In the field of education, statistical analysis plays a crucial role in several aspects. Educational institutions rely on statistical data to evaluate and measure student performance, undertake empirical research on teaching methodologies, and analyze the overall efficacy of educational programs.

Market research is a common practice used by businesses and organizations to get valuable insights pertaining to customer preferences and behavior via the utilization of statistical surveys.

In conclusion, the breadth and application of statistics are extensive, since it offers a theoretical foundation for the acquisition, examination, and comprehension of data across many disciplines. It facilitates the use of evidence-based approaches in decision-making, research, and problem-solving within several disciplines.

1.5 APPLICATIONS OF STATISTICS IN MANAGERIAL DECISIONMAKING

The use of statistics is of utmost importance in the process of management decision-making across a wide range of sectors and functional areas. The use of data analysis, interpretation, and presentation aids managers in making well-informed decisions based on empirical evidence. The following are few prevalent uses of statistics in the context of management decision-making:

Market research is the use of statistical methods to collect and evaluate data pertaining to client preferences, behavior, and market trends. This data facilitates the identification of potential business prospects, the delineation of specific market segments, and the formulation of pricing and marketing strategies.

Financial analysis is the use of statistical methods to assess financial information, including income, costs, and profitability. This facilitates the process of making investment choices, developing budgets, and doing financial forecasts.

Quality control encompasses the use of statistical methodologies, including Six Sigma and control charts, to effectively monitor and enhance the quality of products or services. This results in enhanced

decision-making with respect to process enhancements and allocation of resources.

Inventory management involves the use of statistical inventory models to effectively determine the optimal stock level, with the primary objective of minimizing costs while simultaneously maintaining the availability of products. Managers has the ability to make well-informed judgments pertaining to the processes of ordering, replenishment, and storage.

The use of statistical approaches in the field of Human Resources enables the analysis of employee performance, evaluation of training efficacy, and facilitates decision-making processes pertaining to recruitment, promotions, and remuneration.

In the field of Operations Management, managers use statistical analysis techniques to evaluate production processes and enhance the allocation of resources. This encompasses the process of making decisions about production schedules, the efficient allocation of resources, and the strategic planning of capacity.

Risk management is the use of statistical analysis to evaluate and effectively address prospective risks by examining past data and detecting potential hazards. It facilitates the process of making informed decisions pertaining to insurance, investment, and contingency planning.

The use of Customer Relationship Management (CRM) involves the study of customer data and statistical information to facilitate decision-making processes pertaining to client segmentation, targeted marketing strategies, and the establishment of loyalty programs.

Strategic Planning: Statistics aids in analyzing historical and current data to formulate business strategies. This analysis offers valuable

perspectives on the dynamics of the market, competitive landscape, and potential avenues for development.

Statistical analysis plays a crucial role in informing decision-making processes related to supply chain efficiency, supplier selection, demand forecasting, and logistics optimization within the field of supply chain management.

In the field of project management, statistics play a crucial role in the monitoring of project progress, estimation of project completion deadlines, and effective allocation of resources by managers. This facilitates the process of making prompt judgments and necessary modifications.

A/B testing is a method used in the fields of marketing and online development, which employs statistical analysis to evaluate the efficacy of various techniques, including website design, email campaigns, and ad text.

The evaluation of performance in a corporate context encompasses the use of statistical methods to assess numerous facets, including but not limited to personnel, departments, and goods. This information serves as a guiding factor in making choices on upgrades and the allocation of resources.

The field of environmental management utilizes statistical analytic techniques to inform decision-making processes pertaining to sustainability, pollution control, and adherence to environmental rules.

The field of healthcare management use statistical analysis to enhance patient care, allocate resources efficiently, and facilitate research endeavors. It facilitates the process of clinical decision-making, resource allocation, and quality enhancement.

Through the use of statistical methodologies and tools, managers possess the ability to enhance their decision-making processes by incorporating data-driven approaches. This, in turn, may result in superior results, heightened operational effectiveness, and a distinct competitive edge within their respective sectors.

1.6 LIMITATIONS OF STATISTICS

a. The qualitative aspect has been disregarded.

Statistical tools do not analyze phenomena that cannot be quantitatively described.

These occurrences are not considered within the scope of statistical analysis. These include several aspects like as physical well-being, financial prosperity, cognitive abilities, and so on. The translation of qualitative data into quantitative data is required.

Experiments are now being conducted to quantify the physiological responses of an individual via the collection and analysis of empirical data. In contemporary times, statistics has become an integral component of several domains and worldwide undertakings.

The approach does not address individual entities.

The definition provided by Professor Horace Sacrist elucidates that statistics pertains only to the aggregation of facts or objects, without acknowledging the significance of individual elements. Hence, the isolated occurrences of six fatalities in a single accident and the 85% performance of a certain class in a given year cannot be considered as statistics, since they lack the necessary grouping of related elements. The subject matter does not address the specific elements, regardless of their significance.

The depiction provided is incomplete in capturing the whole of the phenomena.

When various phenomena occur, they are attributed to several causes, although not all of these causes can be quantified or represented in terms of empirical facts. Therefore, it is not possible to arrive at accurate findings. The development of a group is contingent upon several social elements, such as the economic status of parents, education, culture, geography, and government administration. However, it is not possible to include all of these elements in the dataset. We only analyze data via a quantitative lens, disregarding qualitative aspects. The findings or conclusions may not be entirely accurate since they fail to consider several factors.

It has the potential to be misinterpreted.

According to W.I. King, a notable limitation of statistics is their lack of inherent indication of their quality. It might be said that an examination of the data and methodologies used in reaching findings is warranted. However, it is possible that the data in question were gathered by individuals lacking expertise or integrity, leading to potential inaccuracies or biases. Given its intricate nature and susceptibility to unethical manipulation, this field of study has significant implications when wielded by those lacking moral integrity. The use of data necessitates a cautious approach. Alternatively, the outcomes may potentially be catastrophic.

The precision of laws is not absolute.

When considering the two basic rules of statistics:

The principle of the law of big numbers and its application in the field of statistics.

The Law of Statistical Regularity is considered to be less robust compared to scientific laws.

These findings are derived via probabilistic analysis. The outcomes obtained may not consistently match the level of accuracy achieved

by scientific laws. Based on the principles of probability or interpolation, it is possible to provide an estimate of the paddy output in 2008; nevertheless, it is not feasible to assert with certainty that it would precisely amount to 100%. In this context, only estimations or approximations are used.

The findings are valid just in terms of their average representation.

As previously mentioned, the following data have been interpolated, and other methods such as time series analysis, regression analysis, or probability models may be used for further analysis. These statements may not be universally valid. If the average scores of two sections of students in a statistics class are equal, it does not imply that all 50 students in section A have obtained the same marks as those in section B. There is a significant degree of diversity between the two. The outcomes we get are of average kind.

Statistics primarily focuses on the calculation and analysis of averages, which might consist of individual elements that exhibit significant variations from one another. W.L. King is a notable person.

There are several approaches available for studying challenges.

In this field of study, a multitude of methodologies are used to get a singular outcome. Variation may be assessed by several measures such as quartile deviation, mean deviation, or standard deviations, and the outcomes differ across each of these measures.

It should not be assumed that statistics is the exclusive way to use in research, nor should this approach be seen as the optimal strategy for addressing the issue. Croxten and Cowden

1.7 LET US SUM UP

Statistics is an academic discipline that encompasses the systematic investigation of data collection, analysis, interpretation, presentation, and organization. In essence, this pertains to a mathematical field that involves the gathering and summarization of data. Moreover, it may be said that statistics is a discipline within the realm of applied mathematics. Nonetheless, it is important to acknowledge that statistics encompasses two fundamental concepts, namely uncertainty and variation. Statistical analysis is the only means by which one may ascertain the levels of uncertainty and variance across various sectors. These uncertainties are primarily influenced by the probability factor, which has significant importance in the field of statistics.

The use of statistical methods in analysis enables management to get a forecast or an overview of the future market. This approach offers a cost-effective means of making future judgments. Trend analysis, a frequently used statistical tool, examines historical market patterns to make predictions about future trends.

1.8 KEY WORDS

Statistic: is a numerical value that serves as a representation of a certain characteristic or trait of a given sample.

Data : refers to the factual and empirical values that represent the variable under consideration.

Descriptive statistics: pertains to the methodologies and approaches used for the purpose of summarizing and elucidating the attributes and features of the data.

Inferential statistics : pertains to the methodologies that facilitate the process of drawing conclusions or making inferences.

1.9 ANSWERS TO CHECK YOUR PROGRESS

1. Descriptive statistics is a branch of …….

2. The need for inferential statistical methods derives from the need for ……..

3. Numerical facts are usually subjected to statistical analysis with a view to helping a decision maker make wise decisions in the face of ……..

4. Statistics is the science of ……..

5. Statistics deals only with the ……characteristics.

Answer: 1. Statistics, 2. Sampling 3. Uncertainty 4. Numbers 5. Quantitative.

1.10 TERMINAL QUESTIONS

1. Define statistics and discuss its scope.

2. What are the limitations of statistics?

3. Discuss applications of statistics in managerial decision making.

1.11 REFERENCE

1. **Lind, Marchal, Wathen (or Keller)**: *Basic Statistics for Business & Economics* / *Statistical Techniques in Business & Economics*.
2. **Doane, David F.**: *Essential Statistics in Business & Economics*.
3. **Spiegel, Murray R**.: *Statistics* (Schaum's Outline Series) – Great for foundational concepts.
4. **Srivastava & Rego**: *Statistics for Management*
5. **Goon, Gupta & Dasgupta**: *Fundamentals of Statistics* (For deeper theory).
6. **JK Thukral**: *Business Statistics*

UNIT 2: FREQUENCY DISTRIBUTIONS; MEASURES OF CENTRAL TENDENCY: MEAN, MEDIAN, MODE AND THEIR IMPLICATIONS

Structure

2.0 Objectives

2.1 Frequency Distributions

2.2 Introduction of Measures of Central Tendency

2.3 Importance of Measures of Central Tendency

2.4 Requisite for Measures of Central Tendency

2.5. MEAN

2.6 Uses of Mean

2.7 Limitations of Mean

2.8 Calculation of Mean

- 2.8.1. Calculation of Mean in Individual series
- 2.8.2. Calculation of Mean in Discrete series
- 2.8.3. Calculation of Mean in Continuous series

2.9 MEDIAN

2.10 Uses of Median

2.11 Limitations of Median

2.12 Calculation of Median:

- 2.12.1 Calculation of Median in Individual series:
- 2.12.2 Calculation median in discrete series
- 2.12.3 Calculation of median in continuous series:

2.13. MODE

2.14. Uses of mode

2.15. Limitations of Mode

2.16. Calculation of Mode

2.16.1 Calculation of Mode in Individual series

2.16.2. Calculation of Mode in discrete series

2.16.3. Calculation of Mode in continuous series

2.17 Let Us Sum Up

2.18 Key Words

2.19 Answers to Check Your Progress

2.20Terminal Questions

2.0 OBJECTIVES

After studying this unit, you should be able to:

- Understand about frequency distributions
- Describe measures of central tendency
- Describe Applications of Mean, Median and Mode

2.1 FREQUENCY DISTRIBUTIONS

In the field of statistics, frequency distribution refers to the presentation of data that displays the number of occurrences, or frequency, of various values during a certain time period or interval. This presentation might take the form of a list, table, or graphical depiction. There are two distinct categories of frequency distribution, namely grouped and ungrouped. Data refers to a compilation of numerical or value-based information that need proper organization in order to get use from it. In this analysis, we will examine the data and its corresponding frequency distribution.

A frequency distribution is a systematic arrangement that presents the precise count of persons within each category on the measurement scale. A frequency distribution is a statistical representation that gives a comprehensive visual representation of the whole range of scores, presenting the data in a structured manner.

2.2. INTRODUCTION OF MEASURES OF CENTRAL TENDENCY

Measures of central tendency are statistical metrics used to characterize the core or mean value of a given dataset. Measures of central tendency are crucial for comprehending data distributions since they provide valuable insights into the usual or representative value of a dataset. The prevalent metrics of central tendency encompass:

The mean, often known as the average, is calculated by dividing the sum of all values in a dataset by the total number of values. The calculation is performed as follows:

The mean of a set of values may be calculated by dividing the sum of all the values by the total number of values.

The mean is susceptible to the influence of extreme values, rendering it vulnerable to the impact of outliers.

The median is defined as the central number within a dataset that has been arranged in ascending order. In the case when the dataset contains an even number of data points, the median is determined by calculating the arithmetic mean of the two central values. The median exhibits a lower degree of sensitivity to outliers in comparison to the mean.

The mode is the statistical measure that represents the value with the highest frequency of occurrence within a given dataset. A dataset may have a unimodal distribution, characterized by a single mode, or a multimodal distribution, characterized by numerous modes. Alternatively, a dataset may lack a mode if all values occur with equal frequency.

Each of these measures offers distinct perspectives on the central tendency of a dataset, and their selection is contingent upon the characteristics of the data and the particular objectives of the investigation. The following factors should be taken into account:

The mean is used to calculate the arithmetic average; nevertheless, it is important to exercise caution when dealing with outliers, since they have the potential to significantly influence the outcome.

The use of the median as a measure of central tendency is advantageous in scenarios where robustness against outliers is desired.

The mode is a suitable statistical measure for determining the most often occurring value or values within a given dataset. It is particularly useful when seeking to find the most prevalent category or item of interest.

In addition to the aforementioned fundamental measures, there exist supplementary measures such as the weighted mean, which is used when distinct values have varying weights, the geometric mean, which is utilized in the context of geometric data, and the harmonic mean, which finds application in rates and ratios. The selection of an appropriate measure is contingent upon the particular context of the data and the research inquiries one seeks to address.

2.3 IMPORTANCE OF MEASURES OF CENTRAL TENDENCY

Measures of central tendency are statistical metrics that provide significant insights into the core or representative values within a given dataset. Data visualization tools aid in the concise summarization and descriptive representation of data, hence facilitating enhanced comprehension and analysis. The mean, median, and mode are the three fundamental measurements of central tendency. The significance of these factors lies in their inherent importance.

Summary: Central tendency measures are statistical techniques that reduce enormous datasets to a single value or a limited collection of values. This simplification technique proves to be advantageous in promptly comprehending the overarching attributes of the data without necessitating the examination of each individual data point.

The concept of central tendency involves the use of statistical measurements to determine the central or average value of a dataset, serving as a valuable point of reference. This aspect has special significance in scenarios when one seeks to ascertain the usual or representative value of the data.

Comparability refers to the ability to compare distinct datasets, and central tendency measurements provide a standardized approach for comparing their central values. As an example, one may do a comparative analysis of the mean earnings in two distinct urban areas or the median examination results of two separate cohorts.

Predictive Value: Measures of central tendency possess the potential to serve as predictors for forthcoming data points in certain instances. One potential approach for forecasting future sales is to use the average of historical sales data.

Data cleaning is an essential process in the analysis of huge datasets, as it aids in the identification and management of outliers. Central tendency metrics are particularly useful in this regard. The presence of outliers may have a substantial impact on the study, and measures of central tendency aid in the identification of these exceptional data points.

The identification of core values inside data is of utmost importance in diverse decision-making processes. In the field of finance, possessing knowledge about the average return on investment has significant importance when making investment choices.

The comparison of distributions may be facilitated by the use of measures of central tendency, which enable the assessment of the central values of distinct distributions. Understanding the differences between two sets of data is crucial in a variety of scientific and economic contexts.

Graphical Representation: Central tendency measurements are often used in the creation of histograms and other forms of graphical representations for data. They assist in ascertaining the locations of peaks and core tendencies in these visual representations.

Hypothesis Testing: Statistical hypothesis testing use measures of central tendency to assess hypotheses. As an instance, one may conduct a comparison of the average performance between two groups in order to ascertain if there exists a statistically significant difference.

Communication: Central tendency measurements provide a straightforward and readily understandable approach of conveying data to those who may not possess expertise in the subject matter or to a wider audience. The provided information provide a concise and comprehensible overview of the facts.

Although measures of central tendency provide essential insights into data, it is important to acknowledge that they do not give a comprehensive depiction of the whole dataset. In order to have a thorough comprehension of a dataset, it is sometimes imperative to take into account other statistical measures such as measures of dispersion (e.g., variance and standard deviation) as well as the characteristics of the data distribution (e.g., skewness and kurtosis).

2.4 REQUISITE FOR MEASURES OF CENTRAL TENDENCY

Measures of central tendency include statistical metrics that characterize the core or representative value of a given dataset. The mean, median, and mode are widely recognized as the most often used metrics of central tendency. In order to effectively compute these metrics, it is essential to guarantee the fulfillment of certain conditions.

Numerical data, whether real numbers, integers, or ratio data, may be analyzed using measures of central tendency. It is not feasible to compute these metrics for data that is category or nominal in nature.

Understanding the distribution of data is of paramount importance. Is the data set characterized by a unimodal distribution (exhibiting a single peak), a bimodal distribution (displaying two peaks), or a multimodal distribution (manifesting more than two peaks)? The selection of an acceptable measure of central tendency might be influenced by the distribution.

The presence of severe outliers, which are numbers that differ substantially from the remaining data, may greatly distort the mean. In the presence of outliers, the use of the median as a measure of central tendency may be more suitable.

When dealing with continuous data, such as measures like height and weight, it is common to use both the mean and median. In the case of discrete data, such as counts of things, the mode may be a more suitable measure.

Ordinal data refers to data that consists of ordered categories without a set gap between them. When summarizing ordinal data, researchers may choose to use either the median or mode, depending on the specific research topic at hand.

The comparison between symmetric and skewed data reveals that symmetrical data tends to exhibit similarity among the mean, median, and mode, but skewed data may lead to disparities among these statistical measures.

Homogeneity should be maintained in order to maintain consistency within the data being analyzed for a certain group. The combination of data from several populations might result in inaccurate calculations of central tendency.

Interpretation: Select the appropriate measure of central tendency that is most congruent with the research question and effectively conveys the intended message about the data. The mean is a statistical measure that reflects the arithmetic average of a set of values. The median, on the other hand, is a measure that identifies the middle value within a data set. Lastly, the mode is a statistical measure that identifies the value that occurs most often within a given set of data.

When choosing the proper measure of central tendency, it is crucial to take into account the context and qualities of the data. Moreover, the utilization of several measures may provide a more all-encompassing comprehension of the data, particularly in situations where there exist doubts or intricacies inside the information.

2.5 MEAN

In the field of statistics, the concept of "mean" pertains to the arithmetic average of a given collection of numerical values. The measure in question is often used as a means of quantifying central tendency, a statistical concept used to characterize the core or representative value of a given dataset. The sign "μ" is often used to represent the mean of a population, whereas "$\bar{x}$" (pronounced as "x-bar") is typically used to represent the mean of a sample.

The mean is determined by the process of summing all the values inside a given data collection and then dividing the resulting sum by the total count of values. Mathematically, the formula for computing the arithmetic mean is expressed as:

The population mean (μ) refers to the average value of a variable within a population.

The symbol μ represents the population mean, which is calculated by dividing the sum of all individual values (ΣX) by the total number of observations (N).

The sample mean, denoted as $\bar{x}$, is a statistical measure that represents the average value of a set of observations or data points in a sample.

The sample mean, denoted as $\bar{x}$, is calculated by dividing the sum of all observed values (Σx) by the total number of observations (n).

In which location:

The symbol μ represents the population mean.

The symbol $\bar{x}$ represents the sample mean.

The symbol ΣX denotes the summation of all values inside the population.

The symbol Σx denotes the summation of all values inside the sample.

The variable N represents the aggregate quantity of values within the population.

The variable "n" represents the total number of values in the sample.

The mean is a statistical measure that represents the central tendency of a dataset by providing a single numerical number. The aforementioned statistic serves as a valuable tool for characterizing the central tendency or the "mean" value within a given dataset. Nevertheless, the presence of outliers, which are extreme values, might have an impact on it, hence potentially compromising its ability to properly portray the "typical" value, particularly in situations when the data exhibits skewness or contains notable outliers. In such circumstances, other measures of central tendency such as the median or mode may be more suitable.

2.6 USES OF MEAN

The mean, which is often referred to as the average, is a key term in the fields of statistics and mathematics. The measure of central tendency is a statistical concept that offers important insights into a given group of data points. The mean of a dataset is determined by summing all the values inside the dataset and then dividing the sum by the total number of values. The mean is often used in several contexts.

Descriptive statistics often use the mean as a measure to depict the central tendency or average value within a given collection. The provided information conveys a perception of the data's central placement.

Summary of Data: Within the realm of data analysis and reporting, the mean is often used as a means to succinctly summarize data and provide a concise representation of the central trend shown by a given dataset.

Comparison: Mean values may be subjected to comparison across distinct groups or datasets in order to ascertain the presence of statistically significant differences. As an example, one may do a comparison of the average income between two distinct populations.

Forecasting: Within the realm of time series analysis, the use of the mean serves the purpose of establishing a fundamental reference point or pattern in past data. Deviation from the established mean might serve as a means to detect abnormalities or discern patterns.

Imputation refers to the process of filling in missing values in the context of missing data analysis. One often used method for imputation involves using the mean to replace the missing values. For example, in the case when there are missing test results for a cohort of students, it is possible to use the mean score as a means of approximating the absent values.

In the context of manufacturing and quality control, the mean is often used as a metric for monitoring the central tendency of a given process. The presence of deviations from the mean may serve as an indicator of whether the process is under a state of control or not.

Economics: Within the field of economics, the concept of the mean is used for the purpose of computing various indices and measurements, such as the Consumer Price Index (CPI). The CPI serves as a tool for monitoring the average fluctuations in prices pertaining to a certain assortment of products and services.

Survey Data: In the context of survey research, the use of the mean as a statistical measure allows for the comprehension of the central

tendency, namely the average opinion or answer, pertaining to a certain inquiry.

Risk assessment is a crucial aspect in the fields of financial and actuarial sciences. The mean is often used as a statistical measure to approximate anticipated outcomes, such as the projected return on an investment or the projected loss in insurance.

Education: Within the realm of education, the use of the mean as a statistical measure enables instructors to evaluate student performance on assessments and assignments, so facilitating a comprehensive understanding of students' overall academic progress.

Sports: Within the realm of sports, the use of the mean is prevalent in the computation of diverse statistical measures, such as the determination of batting averages in baseball or shooting percentages in basketball.

In the field of medical research, the mean is often used as a statistical measure to summarize patient data, specifically to determine the average age of those participating in a clinical study.

Environmental monitoring is a crucial aspect of environmental research, whereby the mean is used as a statistical measure to monitor and assess average temperatures, pollution levels, and several other environmental parameters.

Psychology: Within the field of psychology, the mean is used as a statistical measure to succinctly describe and facilitate the comparison of data derived from psychological examinations and surveys.

It is essential to acknowledge that while the mean serves as a valuable statistical measure, it might exhibit sensitivity towards outliers, extreme values, or data distributions that are skewed. In

instances of this kind, other measures of central tendency such as the median or mode may be more suitable. Moreover, the mean is but one of many statistical measures that provide valuable insights into data, and its use should be carefully evaluated within the framework of the particular study or analysis being undertaken.

2.7 LIMITATIONS OF MEAN

The mean, which is often referred to as the average, is a widely used measure of central tendency within the field of statistics. Although the statistic in question serves as a significant tool for summarizing data, it is important to acknowledge its inherent limits, since it may not always provide a comprehensive representation of the distribution of the data. The mean has many primary limitations:

The mean is susceptible to outliers, which are numbers that deviate substantially from the remaining data points. The presence of a solitary outlier may have a significant influence on the mean, rendering it an untrustworthy metric in datasets containing outliers.

Lack of Robustness: The mean, as a statistical measure, has a lack of robustness, rendering it susceptible to significant impact from extreme values. On the other hand, it might be argued that robust statistics, such as the median, are less susceptible to the influence of outliers and hence provide a more accurate representation of the central trend in such scenarios.

Skewed distributions are characterized by a lack of symmetry in the distribution of data. In such cases, the mean may not provide an accurate representation of the central tendency or "typical" value. In datasets exhibiting positive skewness, characterized by a rightward tail, the mean exceeds the median, while in datasets displaying negative skewness, characterized by a leftward tail, the mean is lower than the median.

Non-integer data: The calculation of the mean may lack significance when applied to non-integer data. For instance, when considering a discrete variable like as the count of children in a home, the calculated mean value may not always be an integer. This might potentially lead to practical confusion.

Ordinal data refers to a kind of data in which the categories have a certain order, but the gaps between them are not always equal. In such cases, the use of the mean as a metric may lack significance.

The bias of the mean may be influenced by the sample size, exhibiting a tendency to prefer bigger samples. Reduced sample sizes may lead to a diminished level of precision in estimating the mean of the population.

Inapplicability of Mean for Categorical Data: The use of the mean is deemed inappropriate for categorical data due to the absence of numerical values associated with categories, rendering averaging unfeasible. In instances of this kind, other metrics such as the mode (representing the most often occurring category) or proportions may be used.

Information Loss: The use of the mean as the only measure of central tendency has the potential to disguise significant intricacies pertaining to the distribution of data. The process condenses the data into a singular numerical value, which may result in the loss of essential details like the distribution and dispersion of the data.

The mean is contingent upon the scale of the data. The change in units of measurement may have a substantial impact on the mean value, hence posing challenges in the comparison of data across disparate scales.

The presence of non-integer values for the mean might provide challenges when dealing with data that represents discrete quantities.

In order to address these constraints and get a more extensive comprehension of data, it is often advantageous to use other measures of central tendency, such as the median and mode, and to contemplate additional descriptive statistics and visual representations that provide a more full depiction of the distribution of the data.

2.8 CALCULATION OF MEAN

2.8.1. Calculation of Mean in Individual series:

An individual series refers to a specific sort of data series whereby objects or data points are enumerated individually and sequentially. This implies that every element within the series is individually displayed, devoid of any kind of grouping or classification. Individual series are often used as a means of representing data or information that lacks inherent categorization or grouping. In individual series frequencies are not given.

There are two methods for calculation of mean:-

i. Direct Method
ii. Short-cut method

(1). Direct Method

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

Here $\sum X$ = Total of all items of series

N= Total items in series

Example:

Calculate Mean from the following data:

4, 3,6,8,10 15, 20, 17, 12,25

Solution:

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

$$\bar{X} = \frac{4+3+6+8+10+15+20+17+12+25}{10}$$

$$= \frac{120}{10}$$

$$= 12$$

(2). Short-cut method

Here X is calculated using an Assumed Mean and taking deviations from it then apply the following formula :-

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

Here $\sum$ dx = Total of deviation taken from assumed mean

N= Total items in series

Example:

Calculate Mean from the following data:

4, 3,6,8,10 15, 20, 17, 12,26

Solution:

X	Deviation from assumed mean(A=10) dx= X-A
4	-6
3	-7
6	-4
8	-2
10	0
15	5
20	10

17	7
12	2
25	15
	∑dx= 20

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

$$\bar{X} = 10 + \frac{20}{10}$$

$$\bar{X} = 10 + 2$$

$$\bar{X} = 12$$

2.8.2. Calculation of Mean in Discrete series:

In the context of discrete series, the values of variables denote the occurrences or repetitions. This implies that the frequencies are assigned in relation to the various values of the variables. The total number of observations in a discrete series, denoted as N, is equivalent to the summation of the frequencies, represented by the symbol Σf.

There are two methods for calculation of mean:-

i. Direct Method
ii. Short-cut method

(1). Direct Method

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

Here ∑ fx = Total of product of f and x

∑ f or N= Total of frequencies.

Example:

Calculate Mean from the following data:

X	10	20	30	40	50	60	70
f	2	3	5	8	9	7	6

Solution:

X	f	f x
10	2	20
20	3	60
30	5	150
40	8	320
50	9	450
60	7	420
70	6	420
	N= 40	∑f x = 1840

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

$$\bar{X} = \frac{1840}{40}$$

$$\bar{X} = 46$$

(2). Short-cut method

Here X is calculated using an Assumed Mean and taking deviations from it then apply the following formula :-

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

Here ∑ fdx = Total of product of deviation taken from assumed mean and frequencies

∑ f or N= Total of frequencies.

Example:

Calculate Mean from the following data:

X	10	20	30	40	50	60	70
f	2	3	5	8	9	7	6

Solution:

X	f	Deviation from assumed mean(A=40) dx= X-A	f dx
10	2	-30	-60
20	3	-20	-60
30	5	-10	-50
40	8	0	0
50	9	10	90
60	7	20	140
70	6	30	180
	N= 40		∑fdx= 240

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

$$\bar{X} = 40 + \frac{240}{40}$$

$$\bar{X} = 40 + 6$$

$$\bar{X} = 46$$

2.8.3. Calculation of Mean in Continuous series:

In the context of continuous series inside a grouped frequency distribution, the variable's values are organized into many class intervals (e.g., 0-5, 5-10, 10-15), each accompanied by its respective frequencies. The procedure used to calculate the arithmetic mean in a continuous series is similar to that utilized in discrete series. The midpoints of several class intervals replace the class interval in a continuous series. When completed, a continuous series and a discrete series exhibit identical characteristics.

There are two methods for calculation of mean:-

i. Direct Method
ii. Short-cut method

(1). Direct Method

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

Here $\sum$ fx = Total of product of f and x

$\sum$ f or N= Total of frequencies.

Example:

Calculate Mean from the following data:

X	0-5	5-10	10-15	15-20	20-25	25-30	30-35
f	4	6	10	16	18	14	12

Solution:

C.I	M.V.	f	fx

0-5	2.5	4	10
5-10	7.5	6	45
10-15	12.5	10	125
15-20	17.5	16	280
20-25	22.5	18	405
25-30	27.5	14	385
30-35	32.5	12	390
		N= 80	∑fx= 1640

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

$$\bar{X} = \frac{1640}{80}$$

$$\bar{X} = 20.5$$

(2). Short-cut method

Here X is calculated using an Assumed Mean and taking deviations from it then apply the following formula :-

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

Here ∑ fdx = Total of product of deviation taken from assumed mean and frequencies

∑ f or N= Total of frequencies.

Example:

Calculate Mean from the following data:

X	0-5	5-10	10-15	15-20	20-25	25-30	30-35
f	4	6	10	16	18	14	12

Solution:

C.I	X	f	Deviation from assumed mean(A=22.5) dx= X-A	f dx
0-5	2.5	4	-20	-80
5-10	7.5	6	-15	-90
10-15	12.5	10	-10	-100
15-20	17.5	16	-5	-80
20-25	22.5	18	0	0
25-30	27.5	14	5	70
30-35	32.5	12	10	120
		N= 80		$\sum$ f dx= -160

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

$$\bar{X} = 22.5 + \frac{-160}{80}$$

$$\bar{X} = 22.5 - 2$$

$$\bar{X} = 20.5$$

2.9 MEDIAN

The median is a statistical metric used to characterize the center tendency of a given dataset. The median is the central value within a dataset when the values are organized in either ascending or descending order. Put simply, the median is the numerical figure that serves as the dividing point between the upper 50% and lower 50% of a dataset.

In the case of a dataset containing an odd number of values, the median may be determined by identifying the value located in the center of the dataset. In the given dataset [1, 3, 5, 7, 9], the median is determined to be 5 since it represents the central value when the numbers are sorted in ascending order.

In cases when the dataset contains an even number of values, it is customary to compute the median by taking the average of the two central values. In the given dataset [2, 4, 6, 8], the median is calculated as the average of the two middle values, which in this case are 4 and 6. Thus, the median is determined to be 5.

The use of the median is prevalent in statistical and data analysis contexts due to its reduced susceptibility to the influence of severe outliers, in contrast to the mean (average). The use of the median in a dataset is advantageous in situations when outliers or skewed distributions are present, since it offers a more accurate depiction of the "typical" value.

In essence, the median represents the central value within a given dataset, serving as a measure of central tendency that effectively characterizes the central tendency of a data distribution.

2.10 USES OF MEDIAN

The median is a statistical metric that denotes the central value within a dataset when arranged in ascending order. The use of this method in statistics and data analysis is extensive due to its reduced susceptibility to outliers in comparison to the mean. The following are few prevalent applications of the median:

Descriptive statistics include the use of the median as a measure to characterize the central tendency of a given dataset. It provides an indication of the central tendency or average value.

The analysis of data dispersion is essential for gaining insights into the manner in which data is spread or dispersed. When the median closely approximates the mean, it indicates that the data is essentially symmetric. Conversely, if the median greatly deviates from the mean, it suggests the possibility of data skewness.

Addressing Skewed Data: In the context of datasets exhibiting skewed distributions, characterized by a concentration of data points on one side, the median emerges as a more suitable measure of central tendency in contrast to the mean.

The identification of outliers is facilitated by the use of the median, since it is less susceptible to the influence of extreme numbers. Outliers may be defined as data points that exhibit substantial deviation from the median.

When dealing with ordinal data, which refers to categorical data that has a natural order but not necessarily equal intervals, the median is an appropriate measure of central tendency.

The examination of income and salary: The use of median income or median pay is often employed as a means to comprehend income inequalities, as it signifies the income value that is at the midpoint of a certain population. In contrast to the arithmetic mean, which is susceptible to distortion by outliers in the form of very high or low wages, the median emerges as a more resilient metric within this particular framework.

The reporting of median house prices in the real estate industry is favored due to its ability to provide a more accurate depiction of the purchasing power of an average buyer, in contrast to the mean price. The mean price may be influenced by a small number of very high-priced homes, thereby distorting the overall picture.

In the field of healthcare, the median is used as a statistical measure for different assessments, including the determination of median

survival time for patients. This metric serves as an indicator of the point in time at which half of the patient population is alive.

Education: The overall performance of pupils in a class or school may be evaluated by using median test scores or grades.

Sports: Within the realm of sports, the median may serve as a valuable tool for assessing the usual performance of a player, particularly in cases when performance data do not adhere to a normal distribution.

The use of the median in financial markets enables the examination of stock prices, commodity prices, and other financial data in order to comprehend key patterns and probable developments.

In the field of psychology and social sciences, the median is often used as a measure of central tendency for survey answers and test scores, particularly in cases when the data does not exhibit a normal distribution.

In conclusion, the median serves as a significant statistical metric, especially in situations including skewed data, outliers, or ordinal data. This statistical measure offers valuable insights into the core patterns of a dataset, exhibiting more resilience to extreme values compared to the mean.

2.11 LIMITATIONS OF MEDIAN

The median is a valuable statistical metric for characterizing the central tendency of a dataset, especially in cases involving skewed or non-normally distributed data. Nevertheless, it is important to acknowledge the inherent limits associated with this phenomenon.

Insensitivity to Extreme Values: The median exhibits a lower degree of sensitivity towards extreme values, sometimes referred to as outliers, in comparison to the mean. This characteristic of the

median may be seen as both advantageous and limiting. In some instances, it may be necessary to provide additional consideration to outliers, hence rendering the exclusive use of the median insufficient in offering a comprehensive representation of the dataset.

The median is considered to be less exact in conveying information about the distribution of data when compared to the mean. The provided information just indicates the central value or location within the dataset, without taking into account the specific values of individual data points.

The computational complexity of calculating the median is sometimes higher compared to determining the mean, particularly when dealing with extensive datasets. While this issue may not be of great importance for the majority of contemporary computer systems, it is worth noting.

The stability and robustness of the median are contingent upon the size of the sample. In instances when the sample size is limited, it is possible that the median may not effectively capture the central tendency of the dataset.

Continuous distributions have limited applicability when it comes to using the median. The median is often used in a more prevalent manner for data that is either discrete or ordinal in nature. In the context of continuous distributions, it is possible that the lack of an intuitive interpretation may necessitate the use of other measures such as the mean or mode, which may be more suitable.

Insufficient Information Regarding Variability: The median fails to provide any insights into the dispersion or distribution of the data, in contrast to the mean and other metrics such as the standard deviation or range.

Absence of Distinct Mathematical features: Although the mean has distinct mathematical features, such as serving as the equilibrium

point of the dataset, the median does not exhibit comparable properties, rendering it less conducive to certain mathematical procedures.

Dealing with Tied Values: In datasets where values are equal, it is possible for there to exist numerous medians. Although not inherently restrictive, this aspect has the potential to complicate the process of interpretation in some instances.

In conclusion, the median serves as a significant indicator of central tendency, particularly in the context of non-normally distributed data or data sets including outliers. Nevertheless, this approach has several drawbacks pertaining to its accuracy, susceptibility to outliers, and the absence of insights into data variability. It is important to take into account these constraints and choose the suitable metric for one's particular analysis and research goals.

2.12 CALCULATION OF MEDIAN:

2.12.1 Calculation of Median in Individual series:

- Arrange the 'N' measurement in ascending or descending order, as both would give the same answer.
- If N id odd, the median is the middle number.
- If N is even, the median is the mean (average) of the middle two numbers.
- M= Value of $\frac{N+1^{nt}}{2}$ item of the series
- M= Median, N= Number of Items

Example :

- Calculate Median of following data.

  20, 30, 16, 24, 10, 4, 9

Solution:

- Arrange the values in ascending order:

  4, 9, 10, 16, 20, 24, 30

M= Value of $\frac{N+1^{nt}}{2}$ item of the series

M= Value of $\frac{7+1^{nt}}{2}$ item of the series

M= Value of $\frac{8^{nt}}{2}$ item of the series

M= Value of 4^{th} item of the series

value of 4^{th}item is 16

Hence Median is 16

Example:

- Calculate Median of following data.

  20, 30, 16, 24, 10, 4, 9,11

  Solution:

- Arrange the values in ascending order:

  4, 9, 10, 11, 16, 20, 24, 30

M= Value of $\frac{N+1^{nt}}{2}$ item of the series = $\frac{8+1^{nt}}{2}$ item of the series = $\frac{4.5^{nt}}{2}$ item of the series

value of 4.5^{th}item is = $\frac{Value\ of\ 4^{th}\ item\ +Value\ of\ 5^{th}\ item}{2} = \frac{11+16}{2}$

value of 4.5^{th}item is 13.5

Hence Median is 13.5

2.12.2 Calculation median in discrete series:

- Arrange the items in ascending and descending order of their magnitude.
- Compute cumulative frequency (c.f.) by adding their respective frequencies.
- Median no. is located by the following formula:-
- M= Value of $\frac{N+1^{nt}}{2}$ item of the series
- Finally, Median (M) = item corresponding to selected *c.f.*

Example:

x	15	25	32	41	50
f	4	10	8	6	6

Solution:

x	F	c.f.
15	4	4
25	10	14
32	8	22
41	6	28
50	6	34
	N=34	

M= Value of $\frac{N+1^{nt}}{2}$ item of the series

M= Value of $\frac{34+1^{nt}}{2}$ item of the series

M= Value of $\frac{35^{nt}}{2}$ item of the series

M= Value of 17.5^{th} item of the series

value of 17.5^{th}item is 32

Hence Median is 32

2.12.3 Calculation of median in Continuous series:

- Computing median (Continuous series) steps involved are :-
- Arrange the items in ascending order of class-interval.
- Compute cumulative frequency(c.f.) by adding their respective frequencies
- Median No. Is located by the following formula:-
- Median No. or **m** = Value of $\frac{N^{nt}}{2}$ item series
- Finally, we get class-interval, corresponding to selected c.f.

$$\mathbf{M = L_1 + \frac{i}{f} (m-c)}$$

- Where M = Median
- L_1 = Lower limit of selected class interval
- *i* = size of selected interval i.e. (L_2-L_1)
- f = frequency of selected class interval
- c = cumulative frequency of class interval preceding the selected class interval
- m = Median No.

- Note- It should be checked that median will always lie within class interval.

Example:

Class Intervals	0-10	10-20	20-30	30-40	40-50
f	2	5	4	3	6

Solution:

Class Intervals	f	c.f.
0-10	20	20
10-20	50	70
20-30	40	110
30-40	30	140
40-50	60	200
	N= 200	

Median No. or **m** = Value of $\frac{N^{nt}}{2}$ item series

= Value of $\frac{20^{nt}}{2}$ item series

= Value of 10^{th} item of the series

M= L_1 + $\frac{i}{f}$ (m-c)

M= 10.50+ $\frac{5}{6}$ (10- 7)

M= 10.50 + 0.83 x 3

M= 10.50 + 2.49

M= 12.99

2.13. MODE

Within the field of statistics, the mode is a statistical measure used to determine the central tendency of a given dataset. It is defined as the value that occurs most often within the dataset. The mode is the numerical number that exhibits the greatest frequency or frequency density within a particular dataset. The mode is considered one of the three primary measures of central tendency, alongside the mean and median.

The mode may be defined as the value or values that occur most often in a given dataset.

The mode of a given data collection refers to the value or values that exhibit the highest frequency of occurrence. Within a given dataset, the mode is defined as the numerical value that occurs most often. A dataset has the characteristic of being unimodal when it contains a single mode, whereas it is considered multimodal if it contains many modes. Alternatively, a dataset is seen to be devoid of a mode if all values within the dataset occur with equal frequency.

In the given dataset of 2, 4, 4, 7, 7, 7, 9, the mode is determined to be 7 since it exhibits a higher frequency of occurrence (three times) compared to any other value included in the dataset.

It is important to acknowledge that a dataset might exhibit a unimodal distribution, a multimodal distribution (such as bimodal or trimodal), or a distribution devoid of any mode. The presence of numerous modes in a dataset indicates the existence of various values that share the greatest frequency.

2.14. USES OF MODE

The mode is a statistical metric that signifies the value with the highest frequency of occurrence within a given data collection. The metric in question is among the three primary indicators of central tendency, with the mean (sometimes known as the average) and median. The method has many applications across several disciplines and circumstances:

The topic of discussion pertains to descriptive statistics.

The task at hand involves determining the most often occurring or widely accepted value within a given collection of data.

Offering an account of the prevailing or often seen phenomenon.

The process of examining and interpreting data in order to uncover patterns, relationships, and insights.

The process of discerning prevalent trends, patterns, or traits within a given dataset.

Categorical or discrete data, such as various sorts of fruits, automobile colors, or customer feedback scores, may be effectively summarized using this method.

The topic of discussion pertains to quality control.

Within the realm of quality control procedures, the act of determining the manner of faults or difficulties serves the purpose of identifying prevalent problem areas that may be targeted for improvement.

The topic of education is of great significance and relevance in contemporary society.

Within the realm of education, the mode may be effectively used as a statistical measure to ascertain the score that occurs most often on

a given exam. This particular metric serves as an indicator of the overall performance level shown by a collective of students.

The study of market trends and consumer behavior in order to gather information and insights that may be used to make informed business decisions.

This study involves the examination of client preferences, the identification of popular items or services, and the comprehension of market trends by means of analyzing replies obtained from surveys or questionnaires.

The field of healthcare encompasses a wide range of practices, policies, and systems aimed at promoting

The identification of often occurring symptoms or diagnoses in medical data has significant value for illness monitoring, resource allocation, and medical research purposes.

The field of economics is a social science that studies the production, distribution, and consumption.

The objective of this analysis is to determine the modal income level within a certain population or geographic area by examining the distribution of income.

The field of psychology is a scientific discipline that focuses on the study of human behavior and mental

The objective of this inquiry is to discern the behaviors, preferences, or reactions that are often documented in psychological research.

The topic of discussion is transportation.

Examining the prevalent means of transportation used by individuals commuting within a certain region, hence providing valuable insights for urban planning and traffic control strategies.

The field of study that deals with the theory, design, development, and use of computers and computer systems, often known as computer

In the field of data analysis and computer algorithms, the mode is a statistical measure used to identify the element that appears most often in a given list or array. This particular measure has significance in numerous applications, such as code optimization and data management.

The field of study that focuses on the scientific understanding of the environment and its many components, including the interactions between humans and the natural

The objective is to ascertain the mode of certain environmental factors, such as pollution levels, in order to get insight into the prevailing circumstances.

The Intersection between Gaming and Entertainment:

This study aims to ascertain the prevailing gaming genres, movie genres, and entertainment preferences by analyzing user data.

It is essential to acknowledge that the mode may not always be the best suitable measure of central tendency, since it may not sufficiently capture the general properties of the data, particularly in instances when there are many modes or the distribution exhibits significant skewness. The informativeness of the mean and median may vary depending on the context and data distribution.

2.15 LIMITATIONS OF MODE

It seems that you are inquiring about the constraints and shortcomings associated with the notion of "mode." The mode in statistics refers to the value that exhibits the highest frequency of occurrence within a given data collection. Although the mode serves

as a valuable indicator of central trend, it is not without its limitations:

Uniqueness: In contrast to the mean and median, which possess a singular value for a particular data set, it is possible for a data set to exhibit many modes or even lack a mode altogether. In some instances, this phenomenon results in reduced stability and diminished precision.

Sensitivity to minor fluctuations: The mode exhibits a high degree of sensitivity to even little variations in the data. In the event of a modification to a single value within the dataset, there exists the potential for an impact on the determination of the mode, hence potentially compromising its ability to properly reflect the central tendency.

The descriptive capacity of the mode is limited since it just indicates the value that occurs most often within a dataset, without offering insights into the overall distribution of the data. The provided data does not effectively communicate details like the extent of dispersion or the general distribution pattern, unlike the mean and median measures.

Not universally applicable to all sorts of data: The mode is particularly suitable for the analysis of categorical or discrete data, since it allows for the determination of the most often occurring values. Continuous data may have limited usefulness when the values are not clearly identifiable or exhibit a significant level of fluctuation.

Neglects the magnitudes of values: The mode fails to account for the magnitudes of values, hence lacking the ability to discern between big and small values. For instance, if "1" and "100" are both seen as modes inside a dataset, they are considered equally significant.

Lack of robustness in the presence of outliers: The mode is susceptible to strong effect from outliers, hence exhibiting lower robustness compared to the median.

The mode is mostly used as a descriptive statistic and may have limited utility in generating conclusions or doing more sophisticated statistical studies.

In conclusion, while the mode has practical applications, it may not always serve as the optimal measure of central tendency. Consequently, it is imperative to acknowledge its constraints when selecting a statistical summary for one's dataset. Depending on the characteristics of the data and the particular inquiries to be addressed, other metrics such as the mean or median could be more suitable.

2.16 CALCULATION OF MODE:

2.16.1 Calculation of Mode in Individual series:

Example :-

Calculate the mode for the following data:

2,4,7,7,9,9,7,7,10,11,14,20, 45

Solution:

- 2,4,7,7,7,7, 9,9,10,11,14,20, 45
- Only variable 7 repeats itself four times.
- Mode (z)=7

Example :-

Calculate the mode for the following data:

2,4,7,7,9,9,7,10,10,10,11,14,20, 45

Solution:

- 2,4,7,7,7,9,9,10,10,10,11,14,20, 45
- Variable 7 and 10 repeats three times.
- Mode (z)=7 and 10

Example :-

Calculate the mode for the following data:

2,4,5,7,7,9,10,10,11,14,20,20,45,60

Solution:

- 2,4,5,7,7,9,10,10,11,14,20,20,45,60
- Variable 7, 10 and 20 repeats two times.
- Mode (z)=7,10 and 20

2.16.2.Calculation of Mode in discrete series:

By Inspection method

Example: Calculate the mode for the following data:

X	F
10	6
20	8
30	10
40	8
50	20
60	14
70	9

Solution:

- **Variable 50 has maximum frequency, Hence Mode (Z) = 50**

By grouping method:

Value	Frequencies (F)
X	1
2	5
4	12
6	28
8	26
10	33
12	32
14	14
16	12
18	33
20	14

Solution:

Grouping Table						
Value	Frequencies (F)	GROUPING				
X	1	2	3	4	5	6
2	5	17	—	45	—	—
4	12		40		66	—
6	28	54				87
8	26		59	91		
10	33	65			79	
12	32		46			58
14	14	26		59		
16	12		45		59	
18	33	47				—
20	14		—	—		—

Analysis Table										
Column	Size of the items containing Maximum frequency									
	2	4	6	8	10	12	14	16	18	20
1	✓									
2					✓	✓				
3				✓	✓					
4				✓	✓	✓				
5					✓	✓	✓			
6			✓	✓	✓					
Total	1		1	3	5	3	1			

▶ *Variable 10 has highest total five, Hence Mode (Z) = 10*

2.16.3. Calculation of Mode in continuous series:

Example: Calculate the mode for the following data:

C.I.	F
10 — 20	9
20 — 30	15
30 — 40	10
40 — 50	8
50 — 60	20
60 — 70	12
70 — 80	5

▶ **Solution:**

C.I.	F
10 — 20	9
20 — 30	15
30 — 40	10
40 — 50	8 f0
50 — 60	20 f1
60 — 70	12 f2
70 — 80	5

- **Apply formula to calculate the mode.**
- **L1= 50, f0 = 8, f1 = 20, f2 = 12 and i = 60-50 = 10**
- $$Z = L_1 + \frac{f1-f0}{2f1-f0-f2} \times i$$
- $$Z = 50 + \frac{20-8}{2x20-8-12} \times 10 = 50 + \frac{12}{40-20} \times 10$$
- $$= 50 + \frac{12}{40-20} \times 10 = 50 + \frac{12}{20} \times 10 = 50+6$$
- **= 56**
- **Value of Z = 56**

Example:

Calculate mode from following data:

C.I.	F
10 — 20	4
20 — 30	12
30 — 40	40
40 — 50	41
50 — 60	27
60 — 70	13
70 — 80	9
80 — 90	41

Solution:

Grouping Table						
C.I.	Frequencies (F)	GROUPING				
	1	2	3	4	5	6
10 — 20	4	16	—	56	—	—
20 — 30	12		52		93	—
30 — 40	40	81				108
40 — 50	41		68	81		
50 — 60	27	40			49	
60 — 70	13		22			63
70 — 80	9	50		—		
80 — 90	41		—	—	—	

Analysis Table								
Column	Size of the items containing Maximum frequency							
	10 — 20	20 — 30	30 — 40	40 — 50	50 — 60	60 — 70	70 — 80	80 — 90
1				✓				✓
2			✓	✓				
3				✓	✓			
4				✓	✓	✓		
5		✓	✓	✓				
6			✓	✓	✓			
Total		1	3	6	3	1		1

C.I.	F
10 — 20	4
20 — 30	12
30 — 40	40 f0
40 — 50	41 f1
50 — 60	27 f2
60 — 70	13
70 — 80	9
80 — 90	41

- **Apply formula to calculate the mode.**
- **L1= 40, f0 = 40, f1 = 41, f2 = 27 and i = 50-40 = 10**
- $$Z = L_1 + \frac{f1-f0}{2f1-f0-f2} \times i$$
- $$Z = 40 + \frac{41-40}{2x41-40-27} \times 10$$
- $$= 40 + \frac{1}{82-67} \times 10 = 40 + \frac{1}{15} \times 10$$
- $$= 40 + \frac{10}{15} = 40+0.67$$
- **= 40.67**
- **Value of Z = 40.67**

Example: Calculate mode from following data:

C.I.	10 — 20	20 — 30	30 — 40	40 — 50	50 — 60	60 — 70	70 — 80	80 — 90	90 — 100	100 — 110
F	4	6	5	10	20	22	24	6	2	1

Solution:

Grouping Table

C.I.	Frequencies (F)	GROUPING				
	1	2	3	4	5	6
10 — 20	4	10	—		—	—
20 — 30	6		11	15		—
30 — 40	5	15			21	
40 — 50	10		30			35
50 — 60	20	42		52		
60 — 70	22		46		64	
70 — 80	24	30				42
80 — 90	6		8	32		
90 — 100	2				9	—
100 — 110	1	3	—	—		—

Analysis Table										
Column	Size of the items containing Maximum frequency									
	10 — 20	20 — 30	30 — 40	40 — 50	50 — 60	60 — 70	70 — 80	80 — 90	90 — 100	100 — 110
1							✓			
2					✓	✓				
3						✓	✓			
4				✓	✓	✓				

5					✓	✓	✓			
6						✓	✓	✓		
Total				**1**	**3**	**5**	**4**	**1**		

- **Apply formula to calculate the mode.**
- **L1= 60, f0 = 20, f1 = 22, f2 = 24 and i = 70-60 = 10**
- $$Z = L_1 + \frac{f2}{f0+f2} \times i$$
- $$Z = 60 + \frac{24}{20+24} \times 10$$
- $$= 60 + \frac{240}{44}$$
- **= 60 + 5.45**
- **= 65.45**
- **Value of Z = 65.45**

2.17 LET US SUM UP

A measure of central tendency refers to a singular numerical number that seeks to characterize a dataset by pinpointing its center location. Therefore, it is common to refer to measurements of central tendency as measures of central location. Additionally, they are categorized as summary statistics. The mean, also referred to as the average, is a frequently encountered measure of central tendency. However, it is important to note that other measures, such as the median and the mode, also exist.

The mean, median, and mode are all legitimate metrics of central tendency; however, their appropriateness varies depending on the specific circumstances. In the following sections, an examination will be conducted on the concepts of mean, mode, and median, including their calculation methods and the circumstances in which they are most suitable for use.

2.18. KEY WORDS

Central tendency: The concept of central tendency refers to a statistical metric that indicates a singular value encapsulating the whole of a distribution or dataset.

Mean: "mean" is often used to refer to the arithmetic mean of a given dataset.

Median: The Median of any distribution is that value that divides the distribution into two equal parts.

Mode: The mode refers to the value of an observation that has the highest frequency associated with it.

2.19 ANSWERS TO CHECK YOUR PROGRESS

1. The observation which occurs most frequently in a sample is …….

2. Median of the sample 5, 5, 11, 9, 8, 5, 8 is ……..

3. Any measure indicating the centre of a set of data, arranged in an increasing or decreasing order of magnitude, is called a measure of ……

4. The arithmetic mean is highly affected by…….

5. To calculate the median, all the items of a series have to be arranged in a/an ……..

6. The sum of deviations from the _____ is always zero.

Answer: 1. Mode 2. 8 3. Central tendency

4. Extreme values

5. Ascending or descending order 6. **Mean**

2.20 TERMINAL QUESTIONS

Q1. Define median. Also discuss its uses and demerits.

Q1. Define mean. Also discuss its uses and demerits.

Q2. Define mode. Also discuss its uses and demerits.

Q3. Calculate mean, median and mode from the following data ‘

10 ,12, 14, 34, 18, 9, 13, 65, 34,89,25, 3, 4, 3,1, 4, 10

Q4 Calculate mean, median and mode from the following data.

X	3	4	12	20	39	68	70	90	98
F	4	16	7	11	14	12	5	7	9

Q5 Calculate mean, median and mode from the following data.

C.I.	0-10	Oct-20	20-30	30-40	40-50	50-60	60-70	70-80	80-90
F	8	4	14	9	21	19	8	13	10

UNIT 3: MEASURES OF DISPERSION: RANGE, QUARTILE DEVIATION, MEAN DEVIATION AND STANDARD DEVIATION:

Structure

3.0 Introduction

3.1 Objectives

3.2 Characteristics of a Good Measure of Dispersion

3.3 Limitations of measures of dispersion

3.4 Types of Measures of Dispersion

3.5. Range

3.6 Advantages and Disadvantages of Range

3.7 Calculation of Range

3.7.1. Calculation of Range in Individual series

3.7.2. Calculation of Range in Discrete series

3.7.3. Calculation of Range in Continuous series

3.8 Quartile Deviation

3.9 Merits and Demerits of Quartile Deviation

3.10 Calculation of Quartile Deviation

3.10.1. Calculation of Quartile Deviation in Individual series

3.10.2. Calculation of Quartile Deviation in Discrete series

3.10.3. Calculation of Quartile Deviation in Continuous series

3.11 Mean Deviation

3.12 Advantages and Disadvantages of Mean Deviation

3.13 **Uses of Mean Deviation**

3.14 Calculation of Mean Deviation:

3.14.1 Calculation of Mean Deviation in Individual series:

3.14.2 Calculation Mean Deviation in discrete series

3.14.3 Calculation of Mean Deviation in continuous series:

3.15. Standard Deviation

3.16. Properties of standard deviation

3.17. Advantages of standard Deviation

3.18. Disadvantages of standard Deviation

3.19. Uses of Standard deviation

3.20. Calculation of Standard deviation

3.20.1. Calculation of Standard deviation in individual series:

3.20.2. Calculation of Standard deviation in discrete series:

3.20.3. Calculation of Standard deviation in continuous series:

3.21 Let Us Sum Up

3.22Key Words

3.23 nswers to Check Your Progress

3.24 Terminal Questions

3.0 INTRODUCTION

Dispersion in statistics refers to the measure of variability or dispersion shown by a given dataset. Dispersion refers to the phenomenon whereby data becomes distributed, elongated, or diffused among several categories. The process entails determining

the magnitude of distribution values that may be anticipated from the dataset for a given variable. Dispersion in statistics refers to the tendency of numerical data to exhibit variability around the assumption of an average value.

The concept of dispersion in Statistics facilitates comprehension of datasets by categorizing them according to certain measures of dispersion, such as variance, standard deviation, and range.

Dispersion refers to a collection of statistical measurements that enable the objective and measurable assessment of data quality. Data science courses often start by covering fundamental concepts in statistics, and dispersion is an essential subject that should not be overlooked.

The topic of interest pertains to measures of dispersion.

The units of measurement for the measurements of dispersion are nearly identical to those of the corresponding quantities being measured. There are several measures of dispersion that provide a deeper understanding of the data.

The measures of dispersion are basically used to define the spreadness of data around its central point (measures of central tendency).

3.1 OBJECTIVES

After studying this unit, you should be able to:

- Understand about frequency distributions
- Describe measures of central tendency
- Describe Applications of Mean, Median and Mode

3.2 CHARACTERISTICS OF A GOOD MEASURE OF DISPERSION

The calculation process should be straightforward and easily comprehensible.

The analysis should be grounded on a comprehensive examination of all the observations made throughout the series.

The concept in question need to be precisely and strictly defined.

The variable should exhibit resistance to the influence of outliers.

It should not be disproportionately influenced by variations in the sample.

The subject matter should possess the necessary attributes to facilitate further mathematical examination and statistical evaluation.

Advantages of measures of dispersion

Measures of dispersion provide valuable insights into the variability or dispersion of a given group of data points. There are many benefits associated with the use of measures of dispersion.

This study aims to quantify the extent of variability.

Measures of dispersion, such as the range, interquartile range, and standard deviation, are used to quantify the degree to which individual data points depart from the central trend, which may be represented by the mean, median, or mode. This aids in comprehending the general variability present within the dataset.

A Comparative Analysis:

Dispersion measurements facilitate the assessment of variability across distinct data sets. For instance, the comparison of standard

deviations between two sets of data might facilitate the identification of the set that exhibits a greater degree of variability.

Risk assessment is a systematic process used to identify, evaluate, and prioritize potential risks in order to make informed decisions and take appropriate actions

Measures of dispersion play a vital role in the field of finance and risk management as they are essential for evaluating the level of volatility and risk inherent in investment returns. A greater degree of dispersion is indicative of heightened risk, and investors often take this factor into account when formulating investment choices.

The process of assessing the quality of data.

Elevated levels of dispersion might potentially suggest the presence of mistakes, outliers, or inconsistencies within the dataset. The examination of dispersion aids in the identification and exploration of the existence of outliers or atypical patterns within the dataset.

The Importance of Precision in Academic Research:

Measures of dispersion are often used by researchers to articulate the accuracy and dependability of their results. A reduced dispersion implies that the data points exhibit a higher degree of proximity to the center value, hence implying outcomes that are more dependable.

The process of making choices or reaching conclusions based on careful evaluation and analysis of available options and information.

Measures of dispersion are used by decision-makers across several domains to facilitate informed decision-making. In the context of manufacturing, a comprehensive comprehension of the variability inherent in product dimensions may be useful in establishing robust quality control protocols.

The act of predicting or estimating future events or trends based on available information and analysis.

Measures of dispersion are often used in the fields of statistics and economics for the purpose of forecasting. A comprehensive comprehension of variability facilitates the generation of more precise forecasts and the estimation of the probable spectrum of outcomes.

An Evaluation of Diversity

Within the field of social sciences, it is common practice to use measures of dispersion as a means of evaluating the level of variety present within a given group. An illustration of this concept lies in the dispersion of income levels within a given society, which may serve as an indicator of the extent of economic inequality.

The topic of discussion pertains to statistical tests.

Several statistical tests rely on or need certain amounts of variability in the data. Measures of dispersion are used by researchers to verify if their results align with the underlying assumptions of the tests being conducted.

Enhancing the Process of Interpretation:

Understanding the dispersion of data enhances the central tendency measurements, so offering a more comprehensive depiction. The use of this approach serves to mitigate the risk of oversimplifying data, so facilitating a more comprehensive and nuanced analysis of the many attributes inherent in a given distribution.

In conclusion, measures of dispersion are of paramount importance in statistical analysis since they provide significant insights on the extent and arrangement of data. Summary statistics play a crucial role in enhancing the interpretability and utility of information,

hence providing valuable guidance for decision-making across many domains.

3.3 LIMITATIONS OF MEASURES OF DISPERSION

Nevertheless, it is important to acknowledge the inherent limits associated with them.

The analysis conducted in this study demonstrates a high degree of sensitivity to outliers.

The presence of outliers, which are values that deviate greatly from the rest of the data set, may have a substantial effect on measures of dispersion, particularly the range and standard deviation. The presence of a solitary outlier has the potential to significantly alter the overall dispersion of data, hence potentially causing misunderstanding.

The assumption of normality is a fundamental concept in statistical analysis.

Certain measurements of dispersion, such as the variance and standard deviation, are predicated on the assumption that the data adheres to a normal distribution. If the distribution of the data is non-normal, these measurements may not provide an appropriate representation of the variability.

Lacking in robustness:

Measures such as the range and standard deviation are considered to lack robustness as statistical indicators. This implies that the measurements might be susceptible to the effect of outliers, thus distorting the representation of the data's central tendency.

The Potential Impact of Mask Distribution on Shaping Public Health Outcomes

Measures of central tendency serve as summary statistics that provide insights into the general location of values, while measures of dispersion concentrate on the extent of variability within the data. However, it is possible that they do not provide any details on the nature of the distribution, such as its symmetry or skewness.

The concept of unit dependence refers to the phenomenon in which the numerical value of a physical quantity is influenced by the choice of units used

The units of variance and standard deviation correspond to those of the original data, which may provide challenges when comparing the dispersion of datasets with varying units. The use of the coefficient of variation (CV) serves as a measure to mitigate this concern by representing the standard deviation as a proportion of the mean, expressed in percentage terms.

The use of many measures in academic assessment.

Various measurements of dispersion might provide divergent results on the extent of data dispersion. The interquartile range only takes into account the central 50% of the dataset, while the range encompasses all data points. The selection of a suitable metric is contingent upon the unique attributes of the dataset.

Lacking Intuitiveness:

The values of measures of dispersion may not necessarily possess intuitive or readily interpretable characteristics. For example, the use of a standard deviation value of 5 may lack substantial information on the dispersion of the dataset in the absence of appropriate contextualization.

Constrained inside a two-dimensional framework:

Several metrics of dispersion are specifically developed for unidimensional data and may not be immediately applicable to more intricate, multivariate datasets.

In brief, while measures of dispersion provide useful insights into the variability of data, it is essential to exercise caution when interpreting them, taking into account the nature of the data and the unique features of the distribution. Moreover, the use of various measures of dispersion and visual depictions might contribute to a more thorough comprehension of the distribution of data.

3.4 TYPES OF MEASURES OF DISPERSION:

The main Measures of Dispersion are:

1. Range

2. Quartile Deviation

3. Mean Deviation

4. Standard Deviation

3.5 RANGE:

The measure of dispersion known as range is often regarded as the most straightforward to comprehend. The numerical disparity between the maximum and minimum values inside a dataset is often referred to as the range. The user's text may be reformulated to adhere to academic writing standards.

The range (R) is calculated by subtracting the smallest item (S) from the largest item (L).

$$R = L - S$$

The range, which is denoted in the units of measurement of the provided data, serves as an absolute measure of dispersion. A higher range number signifies a higher degree of dispersion, while a lower range value implies a lower degree of dispersion. If all the items in the range are identical, the value of the range will be 0, showing the absence of any variation or dispersion among the objects.

The range is considered an absolute measure of dispersion, and as such, it is not suitable for comparing the variability of two distributions that are expressed in different units. For example, when the measure of dispersion is expressed in monetary units, it is not directly comparable to a measure expressed in units of length such as feet. In circumstances of this kind, it becomes necessary to use a relative measurement, namely the coefficient of range, which has the advantageous quality of being unaffected by the units of measurement.

Coefficient of range

The coefficient of range refers to a statistical measure that quantifies the dispersion or spread of a dataset. It is calculated by dividing the difference between the maximum and minimum values of

The Coefficient of Range refers to the ratio between the difference between the biggest and smallest elements in a distribution and their total. The coefficient of the range may be seen as a relative measure of dispersion.

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

3.6 ADVANTAGES AND DISADVANTAGES OF RANGE

A. Advantages or Applications:

1. The calculation is straightforward and comprehensible, particularly for those who are new to the subject.

2. This is one of the measurements that are well defined in terms of stiffness.

3. By providing a comprehensive overview with a single observation, it offers a holistic understanding of the issue.

4.The purpose of its use is to assess the standard of a product in order to ensure quality control. The consideration of range is a crucial factor in the preparation of R-charts, as it contributes to the maintenance of quality.

5. The consideration of historical price movements is also taken into account while formulating the concept of the relationship between the prices of Gold and Shares.

6.The organization referred to as the "Meteorological Department" is a governmental or institutional body responsible for the study and analysis of weather patterns and atmospheric conditions. Additionally, it predicts weather conditions by monitoring the temperature range.

B. Disadvantages or Constraints or Shortcomings:

1. The determination of range does not rely on the inclusion of all words. The size of an object is accurately represented only by its extreme elements. Therefore, the range may not provide a comprehensive representation of the data, since it disregards any other central values.

2. The lack of reliability in using range as a measure of dispersion is attributed to the aforementioned issue.

3. The range remains constant even when all other intervening terms and variables are altered to the smallest extent.

4. The range is significantly influenced by the variability in sampling. The range exhibits variability among different samples. As the sample size rises, the range also increases, and conversely, as the sample size decreases, the range decreases.

5. This information does not provide any insights on the variability of other data.

6. The range of open-end intervals is considered uncertain due to the absence of lower and upper boundaries for the first and final intervals.

3.7 CALCULATION OF RANGE:

Calculation of range in individual, discrete and continuous series.

3.7.1 Calculation of Range in Individual series:

Example:

Calculate Range and its coefficient from the following data:

24, 34, 54, 75, 18, 29, 33, 87, 27, 65

Solution:

Largest value is 87 and smallest value is 18

$$\text{Range} = L - S$$

$$= 87 - 18$$

$$= 69$$

$$\text{Coefficient of range} = \frac{87-18}{87+18}$$

$$= \frac{69}{105}$$

$$= 0.657$$

3.7.2 Calculation of Range in discrete series:

Example:

Calculate Range and its coefficient from the following data:

X	10	20	30	40	50	60
F	2	4	6	3	5	1

Solution:

Largest value is 60 and smallest value is 10

$$\text{Range} = L - S$$

$$= 60 - 10$$

$$= 50$$

$$\text{Coefficient of range} = \frac{60-10}{60+10}$$

$$= \frac{50}{70}$$

$$= 0.714$$

3.7.3. Calculation of Range in continuous series

Example:

Calculate Range and its coefficient from the following data:

C.I.	10-20	20-30	30-40	40-50	50-60	60-70
F	3	4	5	3	7	9

Solution:

Largest value is 70 and smallest value is 10

$$\text{Range} = L - S$$

$$= 70 - 10$$

= 60

Coefficient of range = $\frac{70-10}{70+10}$

= $\frac{60}{80}$

= 0.75

3.8. QUARTILE DEVIATION (Q.D.) :

The Quartile Deviation, also known as the Semi-Interquartile Range, is defined as half of the difference between the Upper Quartile (Q3) and the Lower Quartile (Q1). In more precise terms, the QD, or quartile deviation, represents one-half of the inter-quartile range. Therefore, the formula used to calculate Quartile Deviation is as follows:

Quartile Deviation: = $\frac{Q3-Q1}{2}$

Where

Q1 = Lower quartile

- Q1= Value of $\frac{N+1^{nt}}{4}$ item of the series

Q3 = Upper quartile

- Q3= Value of $\frac{3(N+1)^{nt}}{4}$ item of the series

The Quartile Deviation is considered an absolute measure of dispersion, making it unsuitable for comparing the variability of several distributions that are stated in different units. Hence, it is essential to ascertain the Coefficient of Quartile Deviation, often referred to as the relative measure of Quartile Deviation, when comparing the variability of several series with disparate units. The

examination of the degree of variance across various series is a subject of study. The formula for calculating the Coefficient of Quartile Deviation is given by:

Coefficient of Quartile Deviation = $\frac{Q3-Q1}{Q3+Q1}$

3.9. MERITS AND DEMERITS OF QUARTILE DEVIATION

The advantages of using quartile deviation are as follows:

1. The calculation is straightforward and the concept is clearly comprehensible.

2. The task does not include significant mathematical complexities.

The use of the middle 50% words makes this metric superior to both the Range and Percentile Range.

The impact of extreme values is mitigated due to the exclusion of the top 25% and bottom 25% of words.

The Quartile Deviation offers a convenient approach for calculating the Standard Deviation using the formula Q.D. = 5M.D. = 4 S.D. This formula allows for a simplified computation of the Standard Deviation.

If we are required to address the middle portion of a series, this is the most appropriate metric to use.

B. Disadvantages or Limitations

The quartile deviation is a statistical measure that quantifies the dispersion or spread of a dataset by calculating the difference between the upper and

Both Q1 and Q3 are positional measurements and hence cannot be subjected to further algebraic manipulation.

The number of calculations performed is extensive; yet, the significance of the resulting outcome is little.

The results are significantly influenced by variations in the samples.

Approximately half of the words do not contribute to the overall outcome, but disregarding the first and last 25% of items may lead to an unreliable outcome.

If the values deviate from regularity, the outcome is significantly impacted.

The term "measure of dispersion" is not applicable in this context, since it does not accurately represent the degree of variability around a central tendency.

The value of the quartile may be same for two or more series, and the quartile deviation is not influenced by the arrangement of words between the first quartile (Q1) and the third quartile (Q3), or beyond these locations.

After carefully evaluating the advantages and disadvantages, it can be inferred that Quartile Deviation should not be unquestioningly depended upon. When dealing with distributions that exhibit a significant degree of fluctuation, it may be seen that the dependability of quartile deviation is diminished.

3.10. CALCULATION OF QUARTILE DEVIATION:

Calculation of Quartile deviation in individual, discrete and continuous series.

3.10.1Calculation of Quartile Deviation in Individual series:

Example:

Calculate Quartile deviation from the following data:

10, 12, 13, 14, 20, 20, 21, 34, 21, 25, 31

Solution:

Arrange data in ascending order

10, 12, 13, 14, 20, 20, 21, 21, 25, 31, 34.

Calculate Value of Q1 and Q3

Q1 = Lower quartile

- Q1= Value of $\frac{N+1^{nt}}{4}$ item of the series
- Q1= Value of $\frac{11+1^{nt}}{4}$ item of the series
- Q1= Value of $\frac{12}{4}$ item of the series
- Q1= Value of 3rd item of the series
- Value of 3rd item of series is 13
- Hence Q1 =13

Q3 = Upper quartile

- Q3= Value of $\frac{3(N+1)^{nt}}{4}$ item of the series
- Q3= Value of $\frac{3(11+1)^{nt}}{4}$ item of the series
- Q3= Value of $\frac{3(12)^{nt}}{4}$ item of the series
- Q3= Value of $\frac{36}{4}$ item of the series
- Q3= Value of 9th item of the series
- Value of 9th item of series is 25
- Hence Q3 = 25

Quartile Deviation: = $\frac{Q3-Q1}{2}$

Quartile Deviation: = $\frac{25-13}{2}$

Quartile Deviation: = $\frac{12}{2}$

Quartile Deviation: = 6

Coefficient of Quartile Deviation = $\frac{Q3-Q1}{Q3+Q1}$

Coefficient of Quartile Deviation = $\frac{25-13}{25+13}$

Coefficient of Quartile Deviation = $\frac{12}{38}$

Coefficient of Quartile Deviation = 0.326

3.10.2 Calculation of Quartile Deviation in discrete series:

Example:

Calculate Quartile deviation from the following data:

X	10	20	30	40	50	60
F	2	4	6	3	5	8

Solution:

Arrange data in ascending order and calculate Cumulative frequency

X	F	CF
10	2	2
20	4	6
30	6	12
40	3	15
50	5	20
60	8	28
	N=28	

Calculate Value of Q1 and Q3

Q1 = Lower quartile

- Q1= Value of $\frac{N+1^{nt}}{4}$ item of the series
- Q1= Value of $\frac{28+1^{nt}}{4}$ item of the series
- Q1= Value of $\frac{29}{4}$ item of the series
- Q1= Value of 7.25th item of the series
- Value of 7.25th item of series is 30
- Hence Q1 =30

Q3 = Upper quartile

- Q3= Value of $\frac{3(N+1)^{nt}}{4}$ item of the series
- Q3= Value of $\frac{3(28+1)^{nt}}{4}$ item of the series
- Q3= Value of $\frac{3(29)^{nt}}{4}$ item of the series
- Q3= Value of 3(7.25) item of the series
- Q3= Value of 21.75th item of the series
- Value of 21.75th item of series is 60
- Hence Q3 = 60

Quartile Deviation: = $\frac{Q3-Q1}{2}$

Quartile Deviation: = $\frac{60-30}{2}$

Quartile Deviation: = $\frac{30}{2}$

Quartile Deviation: = 15

Coefficient of Quartile Deviation = $\frac{Q3-Q1}{Q3+Q1}$

Coefficient of Quartile Deviation = $\frac{60-30}{60+30}$

Coefficient of Quartile Deviation = $\frac{30}{90}$

Coefficient of Quartile Deviation = 0.33

3.10.3. Calculation of Quartile Deviation in continuous series:

Example:

Calculate Quartile deviation from the following data:

Solution:

Arrange data in ascending order and calculate cumulative frequency

C.I.	F	CF
10--20	3	3
20-30	4	7
30-40	8	15
40-50	3	18
50-60	6	24
60-70	7	30
	N= 30	

Apply following formula for calculation of Lower quartile (Q1)

Calculate q1 for quartile class intervals

q1 = Lower quartile

- q1= Value of $\frac{N^{nt}}{4}$ item of the series

$Q1 = L_1 + \frac{i}{f}$ (q1-c)

- Where Q1 = Lower quartile

- L_1 = Lower limit of selected class interval
- i = size of selected interval i.e. (L_2-L_1)
- f = frequency of selected class interval
- c = cumulative frequency of class interval preceding the selected class interval
- q1 = Lower quartile No.
- Note- It should be checked that Lower quartile will always lie within class interval.

Calculate q1 for quartile class intervals

q1 = Lower quartile

- q1= Value of $\frac{N^{nt}}{4}$ item of the series
- q1= Value of $\frac{30^{rt}}{4}$ item of the series
- q1= Value of $\frac{30}{4}$ item of the series
- q1= Value of 7.5^{th} item of the series
- Value of Q1 lies in class interval 30-40
- Here L1 =30, f= 8, q1= 7.5 and c.f. = 7 and i =10
- **Q1 = 30 + $\frac{10}{8}$ (7.5-7)**
- **Q1 = 30 + $\frac{10}{8}$ (0.5)**
- **Q1 = 30 + $\frac{5}{8}$**
- **Q1 = 30 + 0.625**
- **Q1 = 30.625**

Q3 = Upper quartile

Apply following formula for calculation of Lower quartile (Q3)

- Calculate q1 for quartile class intervals
- q3= Value of $\frac{3(N)^{nt}}{4}$ item of the series

Q3 = L_1 + $\frac{i}{f}$ (q3-c)

- Where Q3 = Lower quartile
- L_1 = Lower limit of selected class interval
- *i* = size of selected interval i.e. (L_2-L_1)
- f = frequency of selected class interval
- c = cumulative frequency of class interval preceding the selected class interval
- q3 = Upper quartile No.
- Note- It should be checked that Upper quartile will always lie within class interval.

Calculate q3 for quartile class intervals

- q3= Value of $\frac{3(N)^{nt}}{4}$ item of the series
- q3= Value of $\frac{3(30)^{rt}}{4}$ item of the series
- q3= Value of $\frac{90}{4}$ item of the series
- q3= Value of 22.5^{th} item of the series

- Value of q1 lies in class interval 50-60
- Here L1 =50, f= 6, q3= 22.5 and c.f. = 18

Apply following formula for calculation of Lower quartile (Q3)

- **Q3 = L_1 + $\frac{i}{f}$ (q3-c)**
- **Q3 = 50 + $\frac{10}{6}$ (22.5-18)**
- **Q3 = 50 + $\frac{10}{6}$ (4.5)**
- **Q3 = 50 + $\frac{45}{6}$**
- **Q3 = 50 + 7.5**
- **Q3 = 57.5**

Quartile Deviation: = $\frac{Q3-Q1}{2}$

Quartile Deviation: = $\frac{57.5-30.625}{2}$

Quartile Deviation: = $\frac{26.875}{2}$

Quartile Deviation: = 13.438

Coefficient of Quartile Deviation = $\frac{Q3-Q1}{Q3+Q1}$

Coefficient of Quartile Deviation = $\frac{57.5-30.625}{57.5+30.625}$

Coefficient of Quartile Deviation = $\frac{26.875}{88.125}$

Coefficient of Quartile Deviation = 0.304

3.11. MEAN DEVIATION (M.D.) :

The Mean Deviation of a series refers to the arithmetic average of the departures of different items from a measure of central tendency, such as the mean, median, or mode. The other designations for Mean Deviation include the First Moment of Dispersion and Average Deviation.

The mean deviation is determined by using all of the elements within the series. Theoretical calculation of mean deviation involves the determination of deviations from any of the three available averages. In practice, either the mean or the median is used to ascertain the mean deviation. The consideration of mode is often omitted due to its inherent uncertainty and tendency to provide inaccurate outcomes. The superiority of the sum of deviations from the median over the sum of deviations from the mean arises from the observation that the former is less than the latter.

Please note that while computing deviations from the specified average, the sign (+ or -) of the deviations is disregarded, on the assumption that all deviations are positive.

3.12. ADVANTAGES AND DISADVANTAGES OF MEAN DEVIATION:

Advantages:

The concept is easily comprehensible.

The calculation process is straightforward.

This conclusion is derived from a comprehensive analysis of the whole set of data.

The dispersion, or scatter, of the individual items in a series from its center value is shown.

The impact of extreme values within a series is minimal.

The tool enables the comparison of several objects within a series.

The statement accurately reflects the concept of the arithmetic mean as the measure of central tendency in a dataset.

The practical use of this concept is seen within the realm of business and trade.

Disadvantages

The lack of precise definition in this context refers to the fact that it is calculated based on several central values such as the mean, median, mode, and others, which may lead to varied outcomes.

The algebraic principle is disregarded in the calculation of deviations from the central value of a series, as the + and – signs are ignored.

The subject under consideration does not lend itself to more algebraic analysis.

The observed data is significantly influenced by the variations in sampling.

Determining the precise value of an average whether expressed as a fraction or repeating decimal may be a challenging task. In such instances, using a shortcut technique becomes necessary. This approach entails utilizing a complex formula that necessitates adjustments to accommodate various scenarios.

3.13. USES OF MEAN DEVIATION:

Due of its high level of accuracy and simplicity, economists frequently choose this method.

The calculation of wealth distribution within a society is a valuable tool. This is due to its inclusive nature, including individuals from both the lower and upper ends of the socioeconomic spectrum.

The use of this metric is prevalent in the prediction of business cycles due to its high level of accuracy in capturing variability specifically for this objective.

3.14. CALCULATION OF MEAN DEVIATION:

The calculation of the Mean Deviation involves accumulating the absolute values of the differences between each observation and the average value, and then dividing this sum by the total number of observations.

Calculation of Mean deviation in individual, discrete and continuous series:

When computing the mean deviation for a given individual series, the summation of the deviations from either the mean or median is performed. The summation is then divided by the cardinality of the set.

Mean Deviation: $\frac{\sum|(X-\bar{X})|}{N}$

Coefficient of Mean Deviation: $\frac{Mean\ Deviation}{Mean}$

3.14.1. Calculation of Mean deviation in individual series:

Example :

Calculate Mean deviation and its coefficient from the following data:

10	18	23	15	17	34	29	31	19	21

Solution:

X	Deviation taken from mean $\|(X-\bar{X})\|$
10	11.7
18	3.7
23	1.3

15	6.7
17	4.7
34	12.3
29	7.3
31	9.3
19	2.7
21	0.7
∑X=217	$\sum\|(X-\bar{X})\|$ =60.4

Mean $(\bar{X}) = \frac{\sum X}{N}$

$= \frac{217}{10}$

= 21.7

Mean Deviation = $\frac{\sum|(X-\bar{X})|}{N}$

$= \frac{60.4}{10}$

= 6.04

Coefficient of Mean Deviation = $\frac{Mean\ Deviation}{Mean}$

$= \frac{6.04}{21.7}$

= 0.279

3.14.2. Calculation of Mean deviation in discrete series:

Example:

Calculate Mean deviation and its coefficient from the following data:

X	10	20	30	40	50	60
f	2	4	6	3	5	8

Solution:

X	F	f x	Deviation taken from mean($\|(X-\bar{X})\|$)	$f\|(X-\bar{X})\|$
10	2	20	30.36	60.72
20	4	80	20.36	81.44
30	6	180	10.36	62.16
40	3	120	0.36	1.08
50	5	250	9.64	48.2
60	8	480	19.64	157.12
	N=28	$\sum fX$ = 1130		$\sum f\|(X-\bar{X})\|$ =60.4410.72

Mean $(\bar{X}) = \frac{\sum fX}{N}$

$= \frac{1130}{28}$

$= 40.36$

Mean Deviation = $\frac{\sum f|(X-\bar{X})|}{N}$

$= \frac{410.72}{28}$

$= 14.67$

Coefficient of Mean Deviation = $\frac{Mean\ Deviation}{Mean}$

$$= \frac{14.67}{40.36}$$

$$= 0.363$$

3.14.3. Calculation of Mean deviation in individual, discrete and continuous series:

Example:

C.I.	0-10	10-20	20-30	30-40	40-50	50-60
M.V.	5	15	25	35	45	55

Solution:

C.I.	M. V.	f	f x	Deviation taken from mean($\|(X - \bar{X})\|$)	$f\|(X - \bar{X})\|$
0-10	5	2	10	29.33	58.66
Oct-20	15	4	60	19.33	77.32
20-30	25	8	200	9.33	74.64
30-40	35	3	105	0.67	2.01
40-50	45	6	270	10.67	64.02
50-60	55	7	385	20.67	144.69
		N=30	$\sum fX$ = 1030		$\sum f\|(X - \bar{X})\|$ =421.34

Mean $(\bar{X}) = \frac{\sum fX}{N}$

$= \frac{1030}{30}$

$= 34.33$

Mean Deviation = $\frac{\sum f|(X-\bar{X})|}{N}$

$= \frac{421.34}{30}$

$= 14.04$

Coefficient of Mean Deviation = $\frac{Mean\ Deviation}{Mean}$

$= \frac{14.04}{34.33}$

$= 0.409$

3.15. STANDARD DEVIATION (S.D.) :

Standard Deviation is a commonly used scientific measure of dispersion in statistical analysis for a particular dataset. The alternative term for standard deviation is Root Mean Square Deviation, which derives its name from being the square root of the mean of the squared departures from the arithmetic mean. The Greek symbol σ (sigma) is used to represent the measure of variability known as standard deviation, which was first introduced by Karl Pearson in the year 1893. This approach involves calculating the square root of the arithmetic mean of the squared deviations. The calculation of the deviation of values involves subtracting the arithmetic mean of a given collection of data from each individual value.

According to Spiegel, “The standard deviation is calculated as the square root of the average of the squared differences between each

data point and the mean. The deviations are calculated with respect to the arithmetic mean of the components."

The use of Standard Deviation is often regarded as the most effective method for quantifying the extent of dispersion within a given dataset. The inclusion of both the magnitude and direction of each data point inside a dataset is the reason why standard deviation is used.

Due to its nature as an absolute measure of dispersion, Standard Deviation is not suitable for comparing the variability of two or more series that are given in different units. Hence, it is crucial to ascertain the relative measure of Standard Deviation when comparing the variability of many series with disparate units. Two often used relative measures of Standard Deviation in statistical analysis are the Coefficient of Standard Deviation and the Coefficient of Variation.

The coefficient of standard deviation is a relative measure of dispersion that is calculated by dividing the standard deviation by the mean of a given dataset. The term "Standard Coefficient of Dispersion" is another often used designation for this measure.

3.16. PROPERTIES OF STANDARD DEVIATION

The properties of standard deviation are important to understand in statistical analysis. Standard deviation is a measure of the dispersion or variability of a set of data points. It quantifies how far the data points deviate from the

Several features of standard deviation may be identified.

1.The standard deviation is unaffected by changes in the origin. This principle states that the standard deviation of a sequence of

observations stays unchanged when a constant value is added to or removed from each observation.

2. The computation of the combined standard deviation for two or more groups is analogous to that of the arithmetic mean.

3. In any given collection of data, the standard deviation of the series is always greater than or equal to its mean departure from the mean.

4. The standard deviation is influenced by changes in scale. This implies that when a constant is applied as a multiplier or divisor to all observations, the standard deviation will be correspondingly multiplied or divided by the same constant.

5. The principle that the sum of the squares of deviations of the items from their arithmetic mean is minimized implies that this total is consistently less than the sum of squares of deviations calculated from an imagined mean.

3.17. ADVANTAGES OF STANDARD DEVIATION:

The merits of standard deviation are many and significant. Standard deviation is a statistical measure that quantifies the amount of variability or dispersion in a dataset. It provides valuable insights into the spread of data points around the mean

Standard Deviation has many benefits, which include:

1. The Standard Deviation takes into account every item in a series, considering all values. Hence, each alteration in a single number within the series has an impact on the standard deviation value.

2. Algebraic Approach: Additional algebraic methods may be used when dealing with standard deviation. For instance, given the

knowledge of the standard deviation of distinct groups, it becomes feasible to readily ascertain the combined standard deviation.

3. Enhanced Mathematical Approach: The limitation of disregarding the signs of deviations, as shown in mean deviation, is eliminated in standard deviation by squaring the deviations.

4.The concept of standard deviation is of utmost importance and is extensively used as a measure of dispersion. The measure of dispersion in question is unequivocal and has a precise definition.

5. Minimal Impact of Sampling Fluctuations: When many independent samples are extracted from a same population, it is seen that the standard deviation of the distribution is relatively less influenced by variations between different samples, in comparison to the impact generated by other measures of dispersion.

3.18. DISADVANTAGES OF STANDARD DEVIATION:

There are many drawbacks associated with the use of Standard Deviation:

1. Computational Complexity: In contrast to other measures of dispersion, the computation of standard deviation is characterized by a higher level of difficulty.

2. The standard deviation of a series is contingent upon the units of measurement used for the observations. Hence, it is not appropriate to use standard deviation as a means of comparing the dispersion of distributions that are measured in distinct units.

3. Increased Emphasis on Extreme Values: The standard deviation assigns more significance to extreme values and lesser significance to values that are closer to the mean.

3.19. USES OF STANDARD DEVIATION:

The use of standard deviation in statistical analysis is multifaceted and plays a crucial role in several fields. Standard deviation is a statistical measure that quantifies the amount of dispersion or variability within a dataset.

Standard Deviation may be used to assess and contrast the dispersions of several distributions, provided that the distributions have the same units of measurement and arithmetic means.

The use of Standard Deviation extends to the assessment of the dependability of the mean. In essence, it may be said that a distribution with a smaller standard deviation is seen to have a more dependable.

3.20. CALCULATION OF STANDARD DEVIATION:

The calculation of standard deviation may be performed on three distinct types of series, namely individual, discrete, and frequency distribution or continuous series.

Standard Deviation = $\sqrt{\frac{\sum dx^2}{N} - (\frac{\sum dx}{N})^2}$

3.20.1. Calculation of Standard deviation in individual series:

Standard Deviation = $\sqrt{\frac{\sum dx^2}{N} - (\frac{\sum dx}{N})^2}$

Example:

20	22	23	25	41	34	29	31	27	39

Solution:

X	dx A=25	$\boldsymbol{dx^2}$
20	-5	25
22	-3	9
23	-2	4
25	0	0
41	16	256
34	9	81
29	4	16
31	6	36
27	2	4
39	14	196
	$\sum \boldsymbol{dx}$ = 41	$\sum \boldsymbol{dx^2}$ =627

Mean ($\bar{X}$) = A + $\frac{\sum dX}{N}$

= 25 + $\frac{41}{10}$

= 25 + 4.1

= 29.1

Standard Deviation = $\sqrt{\frac{\sum dx^2}{N} - (\frac{\sum dx}{N})^2}$

$= \sqrt{\frac{627}{10} - \left(\frac{41}{10}\right)^2}$

$= \sqrt{62.7 - (4.1)^2}$

$= \sqrt{62.7 - 16.81}$

$= \sqrt{45.89}$

= 6.78

Coefficient of Standard Deviation = $\frac{Standard\ Deviation}{Mean}$

$= \frac{6.78}{29.1}$

= 0.233

3.20.2. Calculation of Standard deviation in discrete series:

Standard Deviation = $\sqrt{\frac{\sum fdx^2}{N} - (\frac{\sum fdx}{N})^2}$

Example:

X	15	20	34	46	55	60
f	2	4	6	7	5	8

Solution:

X	f	A= 34 dx	fdx	Fdx^2
15	2	-19	-38	722
20	4	-14	-56	784
34	6	0	0	0
46	7	12	84	1008
55	5	21	105	2205
60	8	26	208	5408
	N=32		$\sum fdx$ = =303	$\sum fdx^2$=10127

Mean ($\bar{X}$) = A + $\frac{\sum fdx}{N}$

= 34 + $\frac{303}{30}$

= 34 + 10.1

= 44.1

Standard Deviation = $\sqrt{\frac{\sum fdx^2}{N} - (\frac{\sum fdx}{N})^2}$

= $\sqrt{\frac{10127}{32} - \left(\frac{303}{32}\right)^2}$

= $\sqrt{316.47 - (9.47)^2}$

= $\sqrt{316.47 - 89.68}$

= $\sqrt{226.79}$

= 15.06

Coefficient of Standard Deviation = $\frac{Standard\ Deviation}{Mean}$

$= \frac{15.06}{44.1}$

= 0.341

3.20.3. Calculation of Standard deviation in continuous series:

Standard Deviation = $\sqrt{\frac{\sum fdx^2}{N} - (\frac{\sum fdx}{N})^2}$

Example:

C.I.	10-20	20-30	30-40	40-50	50-60
f	2	4	6	7	5

Solution:

C.I.	M.V.	f	A= 35 dx	fdx	
Oct-20	15	2	-20	-40	800
20-30	25	4	-10	-40	400
30-40	35	6	0	0	0
40-50	45	7	10	70	700

50-60	55	5	20	100	2000
	60	8	25	200	5000
		N=32		$\sum fdx$ = 290	$\sum fdx^2$ = 8900

Mean ($\bar{X}$) = A + $\frac{\sum fdx}{N}$

= 35 + $\frac{290}{32}$

= 35 + 9.06

= 44.06

Standard Deviation = $\sqrt{\frac{\sum fdx^2}{N} - (\frac{\sum fdx}{N})^2}$

= $\sqrt{\frac{8900}{32} - \left(\frac{290}{32}\right)^2}$

= $\sqrt{278.13 - (9.06)^2}$

= $\sqrt{278.13 - 82.08}$

= $\sqrt{196.05}$

= 14.001

Coefficient of Standard Deviation = $\frac{Standard\ Deviation}{Mean}$

= $\frac{15.06}{44.1}$

= 0.341

3.21 LET US SUM UP

Dispersion refers to the process of being scattered or spread out. Statistical dispersion refers to the degree to which numerical data is expected to deviate from a central or average value. In essence,

dispersion serves as a means to comprehend the manner in which data is distributed.

Measures of dispersion are used to elucidate the extent of variability present within a dataset. Dispersion is a statistical concept that pertains to the measurement of the degree to which data points are spread out or dispersed. metrics of dispersion refer to certain sorts of metrics used for quantifying the dispersion of data.

Two data sets may possess identical means, but they might exhibit substantial dissimilarities. In order to effectively explain data, it is essential to possess knowledge on the magnitude of variability. This is determined by the metrics of dispersion. The three often used measures of dispersion in statistical analysis are range, interquartile range, and standard deviation.

3.22 KEY WORDS

Dispersion: Dispersion refers to the process of being scattered or spread out.

Range: Range of a given dataset is defined as the absolute difference between the maximum and minimum values.

Quartile deviation: The quartile deviation is a statistical metric that quantifies the dispersion within the central portion of a dataset.

Mean Deviation: The mean deviation is defined as a statistical measure that is used to calculate the average deviation from the mean value of the given data set.

Standard Deviation: The concept of standard deviation is a statistical metric that quantifies the extent of variation, dispersion, or spread from the mean value.

3.23 ANSWERS TO CHECK YOUR PROGRESS

1. A study in Statistics that helps to interpret the variability of data is known as _____

2. …….. is simply the difference between the maximum and minimum values given in a data set.

3. Range = …….

4.The mean deviation of the data 2, 9, 9, 3, 6, 9, 4 from the mean is…….

5. The numerical value of a standard deviation can never be _____.

6. The quartile deviation of the data 2, 3, 4, 5, 6, 7, 8 is ……

7. + and – sign are ignored in case of …….deviation.

8. If Q3=30 and Q1=10, the coefficient of quartile deviation is…….

9. Standard Deviation is a measure of …….

10. The arithmetic average of the absolute deviation of a series known as the_____

Answer:

1. Measure of dispersion 2. Range 3. 4. 2.57 5. Negative 6. 2

7. Mean 8. 0.5 9. Dispersion 10. Mean Deviation

3.24 TERMINAL QUESTIONS

1 Q. What do you mean by dispersion? Explain its merits and demerits.

2 Q. Define Quartile deviation. Explain its merits and demerits.

3 Q. What do you mean by S.D.? Explain its merits and demerits.

4 Q. Calculate Range, Q.D., M.D. and S.D. from the following data:

X	15	20	25	30	35	40
f	12	24	16	17	28	21

5 Q. Calculate Range, Q.D., M.D. and S.D. from the following data:

34 45 56 31 25 44 21 42

6 Q. Calculate Range, Q.D., M.D. and S.D. from the following data:

C.I.	10-20	20-30	30-40	40-50	50-60	60-70	70-80
f	20	32	31	27	45	26	28

BLOCK II: CORRELATION AND REGRESSION

UNIT 4: MEANING AND USES OF CORRELATION

Structure

4.0 Introduction

4.1 Objectives

4.2 Meaning and definition of correlation

4.3 Correlation and Causation

4.4 Significance of correlation

4.5. Types of Correlation

4.6 Degree of Correlation

4.7 Correlation and causes & effect relationship

4.8 Advantages of Correlation analysis

4.9 Disadvantages of Correlation analysis

4.10Let Us Sum Up

4.11 Key Words

4.12 Answers to Check Your Progress

4.13 Terminal Questions

- Understand linear and non-linear correlation.

4.0 INTRODUCTION:

In several practical scenarios, it is common to encounter situations where data is collected on multiple variables. The below examples will effectively demonstrate the issues.

1. The anthropometric measurements pertaining to the stature and body mass of individuals within a certain cohort. 2. The financial metrics encompassing the income generated from sales activities and the corresponding investment in promotional efforts within a commercial enterprise. 3. The temporal allocation dedicated to educational pursuits and the subsequent academic achievements attained by students during examinations.

A bivariate distribution refers to the availability of data for two variables, denoted as X and Y.

In this correlation analysis, we will examine the case of sales income and advertising spending within the realm of corporate operations. One may inquire if there exists a correlation between sales revenue and advertising expense. What is the relationship between sales income and advertising spend in terms of their impact on each other?

When examining the relationship between the amount of time dedicated to studying and the academic performance of students, a pertinent inquiry arises as to whether there is a correlation between the increase or reduction in study time and the corresponding increase or drop in marks received.

In each of these scenarios, our objective is to establish a connection between two variables, and correlation serves to address the inquiry of whether a link exists between one variable and another.

When there exists a relationship between two variables such that a change in the value of one variable has an impact on the value of another variable, it may be said that the variables are correlated or that there is a correlation between them.

4.1 OBJECTIVES

After studying this unit, you should be able to:

- Define correlation and distinguish it from causation.
- Explain the significance and applications of correlation in various fields.
- Differentiate between positive, negative, and zero correlation.

4.2. MEANING AND DEFINITION:

The term "correlation" is often used in ordinary discourse to indicate a certain kind of relationship or connection. It is worth noting that an association has been seen between days characterized by fog and instances of respiratory distress characterized by wheezing. In statistical terminology, the concept of correlation is used to indicate the relationship between two variables of a quantitative kind. Additionally, we make the assumption that the relationship between the variables is linear, meaning that there is a consistent rise or decrease in one variable for every unit increase or decrease in the other variable. Another often used method in such situations is regression analysis, whereby the objective is to estimate the optimal linear relationship that captures the correlation.

The correlation coefficient, often denoted as r, is a statistical measure that quantifies the strength and direction of the linear relationship between two variables. It is bounded between -1 and +1, inclusive.

A correlation coefficient that is in close proximity to zero, regardless of whether it is positive or negative, indicates a minimal or negligible association between the two variables. A correlation coefficient nearing positive 1 indicates a positive association between the two variables, whereby increments in one variable are linked to increments in the other variable.

A correlation coefficient in close proximity to -1 indicates a negative correlation between two variables, whereby a rise in one variable is linked to a drop in the other one. A correlation coefficient may be computed for variables at the ordinal, interval, or ratio levels of measurement. However, its interpretation becomes limited when applied to variables that are assessed on a nominal scale.

Spearman's rho is used to compute the correlation coefficient for ordinal scales. The correlation coefficient usually used for interval or ratio level scales is Pearson's r, which is sometimes referred to as the correlation coefficient.

In the field of statistics, correlation is a method used to examine and quantify the degree and direction of the association between variables. It is important to note that correlation assesses the level of co-variation between variables, rather than establishing a causal relationship. Hence, it is imperative to refrain from inferring a causal relationship based just on correlation. An illustrative instance can be observed in the presence of a correlation between two variables, X and Y. This correlation indicates that when the value of one variable experiences a change in a particular direction, the value of the other variable is observed to undergo either a concurrent change (i.e., positive change) or a contrasting change (i.e., negative change). Moreover, in the event that a correlation is present, it is of a linear nature, meaning that the relative fluctuations of the two variables may be visually shown by means of a straight line on a graph.

The correlation coefficient, denoted as r, serves as a concise metric that characterizes the degree of statistical association between two variables of interval or ratio level. The correlation coefficient is standardized in such a manner that its values are confined to the range of -1 to +1. When the value of r approaches zero, it indicates a weak correlation between the variables. Conversely, as the absolute value of r increases, in either the positive or negative

direction, the strength of the link between the two variables intensifies.

The two variables are often denoted by the symbols X and Y. To elucidate the relationship between the two variables, X and Y, a scatter diagram is used to visually represent the values of these variables by plotting their respective combinations on a graph. The scatter diagram is first shown, followed by the subsequent presentation of the methodology for calculating Pearson's r. The provided examples consist of rather tiny sample sizes. Subsequently, data obtained from more extensive samples is provided.

As stated by L.R. Connor, the concept of correlation refers to the relationship between two or more variables, where changes in one variable are often followed by matching changes in the others.

According to Croxton and Cowden, correlation is the suitable statistical method for identifying and quantifying a quantitative connection, and representing it concisely in a formula.

As stated by A.M. Tuttle, correlation refers to the examination of the co-variation between two or more variables.

Two variables are said to be correlated when a modification in one variable results in a similar modification in the other variable. To illustrate, A fluctuation in the price of a commodity results in a corresponding alteration in the amount requested. The augmentation of employment levels leads to a corresponding rise in production. As income levels rise, there is a corresponding increase in spending patterns.

The primary focus of analysis in such situations is to the degree of connection among different statistical series.

4.3. CORRELATION AND CAUSATION

The determination of the degree of correlation between two or more variables may be achieved via the use of correlation analysis. Nevertheless, it fails to account for the causal link that exists between variables. If there exists a correlation between two variables, it may be attributed to either of the following factors:

1. Influence of Third Parties: The presence of a third party might lead to a significant level of correlation between the two variables. The present approach fails to include the potential impact of external factors. For instance, there exists a strong association between the yield per acre of grain and jute, since both crops are influenced by the quantity of rainfall. However, in actuality, there is no discernible causal relationship between these two factors.

2. Interdependence (Cause and Effect): When two variables display a strong correlation, it might provide a challenge to ascertain which variable serves as the cause and which serves as the effect. This is due to the potential influence that they may exert on each other. An illustration of this phenomenon may be seen when the price of a commodity rises, resulting in a corresponding increase in its demand. In this context, the price variable is identified as the independent variable, serving as the causal factor, while the demand variable is recognized as the dependent variable, representing the resultant consequence. Nonetheless, it is plausible that the price of the commodity may see an upward trend as a result of heightened demand, which may be attributed to variables such as population growth or other relevant determinants. In this scenario, the cause may be attributed to an escalation in demand, while the consequence manifests as a corresponding change in price.

3. Random Occurrence: There exists the possibility that the observed association between the two variables was only a result of random chance or coincidence. This association is often referred to be spurious. Hence, it is essential to ascertain the potential existence

of a correlation between the variables being examined. For instance, in the absence of any inherent association between the two variables (namely, the income levels of individuals in a given community and their clothing sizes), a notable connection may nevertheless be seen.

It might be argued that correlation just offers a quantitative assessment and does not establish a causal link between variables. Therefore, it is important to guarantee the appropriate selection of variables for the correlation analysis.

4.4. SIGNIFICANCE OF CORRELATION

The significance of correlation lies in its ability to quantify the strength and direction of the relationship between two variables. Correlation coefficients provide a measure of the extent to which changes in

The use of a single figure aids in the determination of the extent of correlation existing between the two variables.

The comprehension of economic behavior is facilitated and essential factors of significance are identified using this approach.

When there is a correlation between two variables, it is possible to make an estimate of one variable based on the value of the other. The aforementioned task is executed using the regression coefficients.

In the realm of business, the concept of correlation plays a crucial role in the decision-making process. Correlation has a crucial role in facilitating predictive analysis, hence contributing to the mitigation of uncertainty. The reliability and proximity to reality of forecasts based on correlation are likely attributable to this phenomenon.

4.5 TYPES OF CORRELATION

Based on direction Correlation can be classified as:

1. Positive Correlation:

A Positive Correlation is seen when two variables exhibit a simultaneous rise or decrease, indicating a direct relationship between them. For instance, the correlation between pricing and supply, income and spending, height and weight, and so on.

2. Negative Correlation:

A Negative Correlation refers to a relationship between two variables in which they exhibit opposing movements, so that a rise in one variable corresponds to a drop in the other variable, and vice versa. One illustrative instance pertains to the correlation between pricing and demand, as well as the connection between temperature and the sale of woollen items, among other factors.

Based on ratio of variations correlation can be classified as:

1. Linear Correlation:

Linear correlation refers to the phenomenon whereby a consistent alteration in the quantity of one variable occurs as a result of a modification in another variable. The phrase "direct variation" is used to describe a situation in which two variables exhibit a proportional relationship, changing in tandem with one another. When two variables that exhibit a consistent ratio of change are shown on a graph, a linear representation will be used to depict the correlation between them. Consequently, it implies a linear correlation.

2. Non-Linear (Curvilinear) Correlation:

A Non-Linear Correlation refers to a situation where there is no consistent alteration in the quantity of one variable as a result of a modification in another variable. The word "non-proportional" is used to describe a situation in which two variables exhibit dissimilar

rates of change. This observation indicates that there is no linear correlation between the variables. For instance, the augmentation of fertilizer use by twofold does not guarantee a proportional improvement in grain yield.

Based on the number of variables involved, correlation can be classified as:

1. Simple Correlation:

The concept of simple correlation pertains to the examination of the relationship between two variables only. An illustration of this may be seen in the correlation between price and demand, as well as the correlation between price and money supply.

2. Partial Correlation:

Partial correlation refers to the examination of the relationship between two variables while controlling for the influence of other factors. The production of wheat is contingent upon a multitude of circumstances, including but not limited to rainfall patterns, the quality of manure used, and the selection of appropriate seeds. However, when examining the association between wheat and seed quality while controlling for consistent levels of rainfall and manure, the observed connection might be seen as incomplete.

3. Multiple Correlation:

Multiple correlation refers to the examination of the relationship between three or more variables concurrently. Simultaneous examination of the whole collection of independent and dependent variables is conducted. An illustrative instance would be the correlation between wheat production and two key factors: seed quality and precipitation levels.

4.6. DEGREE OF CORRELATION

1. Perfect Correlation:

When the connection between two variables exhibits identical proportional variation, it is referred to as perfect correlation. There are two varieties of this phenomenon.

Positive correlation refers to a statistical relationship between two variables in which their proportionate changes occur in the same direction. In this particular instance, the Coefficient of Correlation is represented as a positive value of +1.

Negative correlation refers to a situation where there is an inverse relationship between two variables, so that as one variable increases, the other variable decreases proportionally. In this particular instance, the Coefficient of Correlation is denoted as -1.

2. Zero Correlation:

When there is no discernible association between two series or variables, it is often referred to as having a correlation coefficient of 0 or no correlation. This implies that in the event of a change in one variable without any corresponding effect on the other variable, a lack of correlation between the two variables is seen. In instances of this kind, the Coefficient of Correlation will assume a value of 0.

3. Limited Degree of Correlation:

The current scenario exhibits a restricted level of correlation between perfection and the lack of correlation. Empirical evidence suggests that a restricted level of connection exists in actuality.

In this particular instance, the coefficient of correlation is bounded by the values of +1 and -1.

The correlation coefficient is said to be limitedly negative when there are uneven changes occurring in the opposite direction.

Correlation is considered to be restricted and positive when there are uneven changes in the same direction.

The level of correlation may vary, with a low degree of correlation seen when the coefficient of correlation falls within the range of 0 to 0.25. A moderate degree of correlation is observed when the coefficient of correlation falls within the range of 0.25 to 0.75. Conversely, a high degree of correlation is observed when the coefficient of correlation falls within the range of 0.75 to 1.

4.7 CORRELATION AND CAUSES & EFFECT RELATIONSHIP:

- 1. Both the correlated variables are being affected by a third variable or more than one variable.
- 2. Both the variables might be mutually affecting each other so that neither of them could be designated as a cause of effect.
- 3. Correlation may be due to pure chance.

4.8 ADVANTAGES OF CORRELATION ANALYSIS

Correlation analysis provides numerous benefits in a variety of disciplines, including data analysis, research, and statistics. The following are a few of the primary benefits:

Relationships are identified:

Correlation analysis is a tool that determines the extent to which pairs of variables are correlated. It indicates the intensity (magnitude) and direction (positive or negative) of the relationship.

Ease:

It is simple to calculate and interpret. The majority of statistical software applications offer user-friendly functions for calculating correlation coefficients.

Data Analysis:

In the exploratory data analysis phase, it is a valuable instrument for identifying potential relationships between variables prior to the execution of more intricate analyses.

Predictive Power:

Although correlation does not necessarily imply causation, a robust correlation may suggest a promising candidate for predictive modeling. In regression models, variables that are significantly correlated with an outcome variable can serve as valuable predictors.

Hypothesis Testing:

It serves as a foundation for the evaluation of hypotheses regarding the relationships between variables. For instance, researchers may evaluate whether the correlation between two variables is substantially distinct from zero.

Complexity Reduction:

By identifying groups of variables that move together, correlation can assist in reducing the dimensionality of data in multivariate analysis.

Multicollinearity Identification:

Correlation analysis can assist in the identification of multicollinearity among predictor variables in regression analysis, which can impact the interpretability and stability of regression coefficients.

Basis for Additional Investigation:

It serves as a foundation for more sophisticated statistical methods, including factor analysis, principal component analysis, and structural equation modeling, which frequently depend on comprehending the correlations between variables..

Visualization:

Scatter diagrams or correlation matrices can be employed to visualize correlation analysis, thereby simplifying the communication of results to individuals who lack a statistical background.

Non-parametric alternatives:

There are non-parametric methods for correlation analysis, such as Spearman's rank correlation, that are flexible for a variety of data types and do not necessitate the assumption of normally distributed data.

Outlier Identification:

Analysts can identify outliers that may impact the analysis and interpret the data accordingly by analyzing the scatter plot and correlation coefficient.

In general, correlation analysis is a fundamental and versatile instrument in the data analyst's arsenal, offering critical insights into the relationships between variables.

4.9 DISADVANTAGES OF CORRELATION ANALYSIS

Although correlation analysis is beneficial, it has numerous drawbacks:

Causality Confusion: Correlation does not necessarily indicate causation. The mere fact that two variables are correlated does not imply that one is the cause of the other.

Omitted Variable Bias: The presence of an omitted variable that affects both variables under consideration can influence the correlation, resulting in a spurious relationship.

Non-Linearity: Correlation quantifies the extent of a linear relationship between variables. Correlation analysis may be deceptive if the relationship is non-linear.

Sensitivity to Outliers: Outliers can have a substantial impact on correlation coefficients, thereby distorting the true relationship between the variables.

Only Two Variables Are Considered: Standard correlation analysis is restricted to pairings of variables, which may result in the loss of complex interactions among multiple variables.

Scale Dependence: The correlation coefficient does not provide information regarding the relationship's magnitude; rather, it indicates its direction and strength.

Homoscedasticity is assumed by correlation, which implies that the variability of one variable remains constant at all levels of the other variable. This may not always be the case.

Sampling Errors: The correlation coefficient may be highly variable and unreliable in small samples, which may result in inaccurate conclusions.

Complexity of Interpretation: Correlation values that are close to zero can be challenging to interpret, as they may suggest the absence of a relationship, a non-linear relationship, or the presence of confounding factors.

It is essential to comprehend these constraints in order to interpret correlation results and derive conclusions from them.

4.10 LET US SUM UP

Correlation analysis is a statistical technique used to assess the magnitude and direction of the linear association between two variables. Below are few essential aspects of correlation analysis:

The correlation coefficient (r) is a statistical measure that quantifies the strength and direction of the linear relationship between two variables.

Quantifies the magnitude and orientation of a linear correlation between two variables.

The values span from -1 to 1.

r = 1 indicates a perfect positive connection.

When r = -1, it indicates a perfect negative correlation.

r = 0 indicates no linear association.

Correlation may be classified into many types:

Positive correlation refers to a situation where a rise in one variable is accompanied by a corresponding increase in another one.

Negative correlation refers to a situation when a rise in one variable is accompanied by a drop in another one.

Lack of Correlation: There is no observable relationship between the fluctuations of variables.

Computing Correlation:

Pearson correlation Coefficient: Quantifies the degree of linearity in connections.

Spearman Rank correlation is a statistical measure that quantifies the strength and direction of a monotonic link between two variables, without assuming a linear relationship.

Constraints:

Correlation may not always indicate causality.

Prone to being influenced by extreme values.

Only assesses the strength and direction of linear associations (Pearson correlation).

Graphical depiction:

Scatter plots are often used to depict the correlation between variables.

4.11 KEY WORDS :

Correlation analysis: is a statistical technique used to assess the magnitude and direction of the linear association between two variables.

Correlation coefficient (r) : is a statistical measure that quantifies the strength and direction of the linear relationship between two variables.

Perfect correlation: connection between two variables exhibits identical proportional variation,

4.12 ANSWERS TO CHECK YOUR PROGRESS

1. Correlation analysis is a.....
2. If a modification in one variable leads to a comparable modification in another variable, then the variables are said to be.....
3. When the values of two variables move in the same direction, correlation is said to be .....correlation.
4. Non-linear correlation is sometimes referred to as.....
5. The coefficient of correlation quantifies the relationship between.....
6. The coefficient of correlation is bounded between.....
7. If r= +1, the correlation is considered to be.....

8. The rank correlation coefficient was first found by.....

9. The coefficient of concurrent deviation is dependent on the signs of the.....

10. The term used to describe the investigation of the correlation between two specific sets of data is known as..... correlation

Answer:

1. Multivariate and Bivariate analysis
2. Correlated
3. Positive
4. Curvy linear
5. Variables
6. +1 to - 1
7. Perfect positive
8. Spearman
9. Deviations
10. Simple

4.13 TERMINAL QUESTIONS:

Q.1. What do you mean by correlation? Explain types of Correlation?

Q2. What are the ssignificance of correlation?

Q3. Explain types and degree of correlation.

Q.4. Explain methods of determining correlation.

UNIT 5: MEANING AND USES OF REGRESSION

Structure

5.0 Objectives

5.1 Introduction

5.2 Objectives of Regression Analysis

5.3 Scope of Regression analysis

5.4 Benefits of Regression Analysis

5.5 Significance of regression analysis as a decision-making tool is as follows:

5.6 Analysis of Regression by Scope

5.7 Types of Regression Analysis:

5.8 Uses of Regression analysis:

5.9 Difference between Correlation and Regression

5.10 Let Us Sum Up

5.11 Key Words

5.12 Answers to Check Your Progress

5.13 Terminal Questions

5.0 OBJECTIVES

After studying this unit, you should be able to:

- Define regression and distinguish it from correlation.
- Explain the importance and applications of regression analysis in various fields.
- Understand types of Regression Analysis.
- Describe objectives of Regression Analysis

5.1 INTRODUCTION

Regression analysis is a robust statistical technique that enables the examination of the association between two or more variables of interest. Various forms of regression analysis exist, but fundamentally, they all investigate the impact of one or more independent variables on a dependent variable The use of this method allows for the assessment of the strength of the link between variables, as well as the modeling of their future association.

Regression analysis encompasses several forms, including linear regression, multiple linear regression, and nonlinear regression. The two most often encountered models in statistical analysis are the basic linear model and the multiple linear model. Nonlinear regression analysis is often used to analyze complex data sets characterized by a nonlinear association between the dependent and independent variables.

Regression is a statistical technique often used in the fields of finance, investing, and several other disciplines. Its primary objective is to assess the magnitude and nature of the association between a single dependent variable (typically represented as Y) and a set of independent variables.

Linear regression, often known as simple regression or ordinary least squares (OLS), is a widely used statistical approach. Linear regression is a statistical technique used to model the linear association between two variables by determining the line of greatest fit. Linear regression is visually represented by a straight line, where the slope of the line indicates the extent to which a change in one variable influences a change in the other. The y-intercept in a linear regression model signifies the value of one variable when the other variable is equal to zero. Non-linear

regression models are also present in the field, however characterized by a somewhat higher level of complexity.

Regression analysis is a robust methodology used to elucidate the relationships between observable variables in data. However, it is important to note that regression analysis does not readily establish causality. The term is used in several situations within the realms of business, finance, and economics. For example, this tool is used to assist investment managers in the process of asset valuation and in comprehending the interconnections between various aspects, such as the prices of commodities and the stocks of enterprises engaged in the trading of those commodities.

The purpose of regression analysis is to determine the association between variables, allowing for the estimation or prediction of the value of one variable based on the known value of another variable.

- According to M.M. Blair, "Regression analysis is a mathematical measure of the average relationship between two or more variables in terms of the original unit of data."
- According to Wallies and Roberts, "It is often more important to find out what the relation actually is in order to estimate or predict one variable(the dependent variable) and the statistical technique appropriate to such a case is called regression analysis."

The variable that you are attempting to predict or explain is known as the dependent variable (Y), also referred to as the outcome variable or response variable.

Independent Variable (X): These variables, which are also referred to as explanatory variables or predictor variables, are employed to forecast the value of the dependent variable.

Simple Linear Regression: This regression analysis method approximates the relationship between two variables by fitting a linear equation to the observed data. The equation of the line is typically expressed as $Y = a + b X$ Y=a+bX, where Y Y is the dependent variable, X X is the independent variable, a a is the intercept, and b b is the slope.

Multiple Linear Regression: This method augments simple linear regression by employing multiple independent variables to forecast the dependent variable. The equation is typically expressed as $Y = a + b 1 X 1 + b 2 X 2 + \ldots + b n X n$

Y = a + b 1 X 1 + b 2 X 2 + … + b n X n .

Coefficient (b): In the regression equation, coefficients denote the change in the dependent variable that corresponds to a one-unit change in the independent variable. The slope of the line is denoted by b in simple linear regression. Each independent variable has its own coefficient in multiple linear regression.

Intercept (a): The intercept is the anticipated value of the dependent variable when all independent variables are equal to zero. It is the point at which the regression line intersects the Y-axis.

Residuals are the discrepancies between the predicted values of the regression model and the observed values of the dependent variable. They are employed to evaluate the model's accuracy.

R-squared (R^2) is a statistical measure that quantifies the extent to which the independent variables in the model account for the variance of the dependent variable. The model's fit is better indicated by higher values, which span the range of 0 to 1.

P-value: The p-value is a metric that is used in regression analysis to ascertain the significance of the results. A statistically significant

relationship between the independent and dependent variables is indicated by a low p-value (typically ≤ 0.05).

Premises of Regression Analysis:

Linearity: The independent and dependent variables exhibit a linear relationship.

Independence: Observations are not dependent on one another.

Homoscedasticity: The residuals exhibit a constant variance at all levels of X X.

Normality: The residuals of the model are distributed ordinarily.

5.2 OBJECTIVES OF REGRESSION ANALYSIS

Comprehension of Interrelationships Between Variables:

The primary goal of regression analysis is to investigate and quantify the relationship between a dependent variable and one or more independent variables. For example, regression may be implemented in economics to ascertain the extent to which fluctuations in interest rates (independent variable) influence consumer expenditure (dependent variable).

Prediction and Forecasting: Regression models are frequently employed for predictive purposes. One can create a model that anticipates future outcomes by examining historical data. For instance, regression analysis may be implemented by businesses to anticipate sales by considering seasonality, advertising expenditure, and other variables.

Identifying Key Influencers: In numerous situations, it is essential to determine the factors that have a substantial impact on the outcome of interest. Regression analysis facilitates the identification

of these critical influencers, thereby facilitating targeted interventions. For example, in the field of public health, the identification of the lifestyle factors that have the greatest impact on the risk of cardiac disease can serve as a foundation for the development of public health policies.

Causality Assessment: Regression analysis is primarily used to identify correlation; however, it can also offer insights into potential causal relationships. Regression can assist in the inference of causality when implemented in conjunction with experimental or longitudinal data. However, this frequently necessitates meticulous evaluation of assumptions and confounding factors.

Process Optimization: Organizations can optimize processes and resource allocation by comprehending the relationship between variables. For instance, regression analysis may be implemented by a manufacturing organization to optimize production processes by evaluating the influence of diverse input variables on output quality and quantity.

5.3 SCOPE OF REGRESSION ANALYSIS

Both small and big enterprises are inundated with a vast quantity of data. Regression analysis is increasingly being embraced by individuals and organizations as a means to enhance decision-making processes and mitigate reliance on conjecture. This methodological approach to management is valued for its scientific underpinnings, which contribute to more informed and evidence-based decision-making.

Regression analysis is a statistical technique used by experts to examine and assess the association between different variables, enabling them to make predictions about the future attributes of this connection.

Organizations have the ability to use regression analysis in several ways. A few of them:

Regression analysis is a commonly used method by financial professionals to predict and evaluate potential future possibilities and dangers. The Capital Asset Pricing Model (CAPM) is a widely used regression model in the field of finance that aims to determine the connection between the anticipated return of an asset and the corresponding market risk premium. It serves as a valuable tool for asset valuation and the identification of capital expenditures. Regression analysis is used for the purpose of computing beta (β), which is defined as the measure of return volatility in relation to the broader market for a particular company.

Regression analysis is used by insurance companies to predict the creditworthiness of individuals who have insurance policies. Additionally, it might aid in determining the quantity of claims that can be made within a certain timeframe.

Sales forecasting use regression analysis as a statistical method to anticipate future sales by examining historical data and performance. This process may provide individuals an understanding of past successes, the resultant effects they have generated, and potential areas for improvement in order to yield more precise and advantageous outcomes in the future.

Regression models have been identified as a crucial tool for optimizing corporate operations. In contemporary business practices, managers widely acknowledge the significance of regression analysis as an essential tool for identifying key factors that significantly influence operational efficiency and revenue generation. By using regression analysis, managers may get novel insights and rectify process flaws, so enhancing overall organizational performance.

5.4 BENEFITS OF REGRESSION ANALYSIS

In the practical realm, a multitude of elements contribute to the growth trajectory of a corporation. Frequently, these elements exhibit interdependence, whereby a modification in one aspect might yield either advantageous or detrimental consequences for the other.

There are two main advantages associated with the use of regression analysis to assess the impact of variable modifications on company operations.

Businesses employ regression analysis as a means of making data-driven decisions in their future planning endeavors. This analytical technique aids in identifying the variables that exert the most substantial influence on the outcome, based on prior outcomes. Organizations may enhance their ability to prioritize effectively by using forecasting techniques and relying on data-driven projections.

Identifying avenues for enhancement: Regression analysis, being a method that elucidates the associations between two variables, may be used by firms to discern potential areas for development in terms of personnel, strategies, or tools via the observation of their interplay. An illustration of this concept is that augmenting the workforce on a project might potentially have a favorable effect on the expansion of income.

5.5 SIGNIFICANCE OF REGRESSION ANALYSIS AS A DECISION-MAKING TOOL IS AS FOLLOWS:

Regression analysis is an effective instrument for decision-making. It informs strategies and actions by providing a clear comprehension

of the relationships between variables. For example, regression analysis may be implemented by policymakers to assess the potential influence of policy modifications on economic indicators.

Risk Management: Regression models are employed in the financial and insurance sectors to evaluate and mitigate risk. Companies can mitigate potential losses by devising strategies that consider the impact of a variety of factors on financial outcomes. For instance, regression analysis can assist in the pricing of insurance policies by evaluating risk factors.

Scientific Research: Regression analysis is a fundamental component of scientific research. It enables researchers to evaluate hypotheses and quantify the relationships between variables, thereby advancing knowledge in a variety of fields. For example, regression models are beneficial in the medical field as they facilitate comprehension of the impact of treatments on health outcomes.

Business Optimization: Regression analysis is employed by companies to enhance their marketing strategies and operations. Businesses can enhance their efficiency and profitability by making informed decisions based on data on sales, customer behavior, and market trends. For example, regression analysis can assist in the identification of the most effective pricing strategy for a new product.

Quality Control: Regression analysis is employed in the fields of engineering and manufacturing to enhance quality control. By modeling the relationship between process variables and product quality, companies can identify the factors that contribute to defects and implement adequate corrective measures. This results in decreased expenses and improved product quality.

Education and Human Resources: Regression analysis is employed by educational institutions and HR departments to comprehend the factors that influence performance and outcomes. Schools, for instance, may investigate the correlation between employee productivity and training programs, while HR departments may investigate the influence of class size and instructional strategies on student achievement.

5.6 ANALYSIS OF REGRESSION BY SCOPE

Non-linear vs. Linear Models:

A diverse array of models is included in regression analysis. The most fundamental is linear regression, which presupposes a linear relationship between the dependent and independent variables. Nevertheless, the development and application of non-linear regression models are frequently necessitated by the non-linearity of relationships in real-world data.

Multiple Regression vs. Simple Regression: Multiple regression utilizes two or more independent variables, while simple regression utilizes only one. Multiple regression models can offer a more thorough comprehension of the factors that affect the dependent variable; however, they necessitate additional data and meticulous interpretation to prevent overfitting.

Quantitative vs. Qualitative Variables: Regression is typically used to analyze quantitative data; however, it can also incorporate qualitative variables by employing methods such as dummy variable coding. This broadens the scope of regression analysis to encompass categorical data, including gender, region, and product type.

Time Series Analysis: Regression models can be employed to analyze time series data, in which the dependent variable is influenced by time. Autoregressive integrated moving average

(ARIMA) and other time series regression models are employed to forecast and analyze time-dependent data.

Hierarchical and Multilevel Models: In sectors such as healthcare and education, data frequently exhibits a hierarchical structure (e.g., patients within hospitals, students within schools). These nested data structures are accounted for by hierarchical and multilevel regression models, which enable more precise and nuanced analyses.

Robust and Regularized Regression: Traditional regression methods may not perform well in the presence of outliers or multicollinearity. Regularized regression methods, such as Lasso and Ridge regression, address multicollinearity by incorporating penalty terms into the regression equation, whereas robust regression techniques reduce the impact of outliers.

5.7 TYPES OF REGRESSION ANALYSIS:

1. Simple regression:

The utilization of simple linear regression is used to evaluate the association between two variables of a quantitative kind. Simple linear regression is a suitable statistical technique to use when one seeks to ascertain:

The strength of the association between two variables, such as rainfall and soil erosion, is a significant aspect of study.

The dependent variable's value at a certain value of the independent variable, such as the quantity of soil erosion at a particular degree of rainfall.

2. Multiple regression:

Multiple linear regression (MLR), commonly referred to as multiple regression, is a statistical methodology that use numerous explanatory factors to forecast the result of a response variable. The objective of multiple linear regression is to establish a mathematical model that represents the linear association between the explanatory (independent) factors and the response (dependent) variables. numerous regression may be seen as an expansion of ordinary least-squares (OLS) regression, since it encompasses the inclusion of numerous explanatory variables.

3. Linear regression:

Linear regression is a statistical technique that aims to establish the association between two variables by using a linear equation to model observed data. One of the variables is intended to serve as the independent variable, while the other is intended to function as the dependent variable. An illustration of a linear relationship may be seen between an individual's weight and their height. Therefore, this demonstrates a direct correlation between an individual's height and weight. As the vertical dimension of an individual's body increases, there is a corresponding rise in their body mass.

The presence of a dependent variable or a causal link between variables is not always a requirement; nonetheless, a significant association between the two variables exists. In instances of this kind, a scatter plot is used to infer the magnitude of the association between the variables. In the absence of any discernible association or correlation between the variables, the scatter plot fails to demonstrate any discernible trend of either increase or decrease. In the context of the provided data, the linear regression model does not provide advantageous results.

Regression analysis is a statistical tool that is both adaptable and powerful, with a diverse array of applications in a variety of fields. Its primary objectives include the following: the identification of key influencers, the evaluation of causality, the optimization of processes, and the comprehension of relationships between variables. Additionally, it includes predictive and forecasting. Regression analysis encompasses a diverse array of techniques and models, specifically designed to address a wide range of data types and research queries.

Regression analysis is essential. It is indispensable in numerous disciplines, including business optimization, scientific research, risk management, decision-making, and quality control. By proposing a rigorous methodology for data analysis and the identification of noteworthy patterns, regression analysis enables organizations and individuals to make well-informed, data-driven decisions.

In an era where data is abundant and essential for success, the ability to effectively employ regression analysis is a valuable skill. The scope and applications of regression analysis are anticipated to expand further as technology and methodologies continue to evolve, thereby providing new insights and opportunities for innovation and development across a variety of domains.

4. Curvi - linear or non - linear regression

The term "curvilinear regression" refers to a regression model that aims to match a curve rather than a linear relationship. Frequently seen instances of curvilinear regression models encompass: Quadratic regression is used in cases when there is a quadratic association between a predictor variable and a response variable.

5.8 USES OF REGRESSION ANALYSIS:

The following are some of the most prevalent applications of regression analysis:

Forecasting and Prediction: Regression is a widely used method for forecasting and predicting. Regression models are employed by businesses to forecast sales, demand, or stock prices by analyzing a variety of factors, such as historical data, economic indicators, or advertising expenditures.

Relationship Identification: It assists in the identification and quantification of the strength of relationships between variables. For example, regression can be employed in the healthcare sector to ascertain the impact of various lifestyle factors on blood pressure.

Trend Analysis: Regression analysis is capable of detecting trends that persist over time. For instance, it may be implemented to evaluate temperature fluctuations over the course of several decades in order to comprehend the progression of climate change.

Error Reduction: Regression analysis reduces errors in predictions by modeling the relationship between variables. This is especially beneficial for the enhancement of operational efficiency and quality control.

Risk Management: Financial institutions employ regression modeling to evaluate risk by examining the correlation between investment returns and a variety of economic indicators.

Regression analysis is implemented by organizations to optimize processes and outcomes. For instance, it can assist in identifying the most profitable price point.

Evaluation of Policies and Interventions: Regression analysis is employed by governments and organizations to assess the

effectiveness of policies or interventions. For instance, it may evaluate the efficacy of a novel educational initiative with respect to student achievement.

Marketing Research: Regression is employed by marketers to ascertain the influence of various marketing initiatives on sales, including the influence of social media campaigns on consumer purchasing behavior.

Scientific Research: In scientific studies, regression analysis is used to comprehend the relationship between variables, such as the impact of a substance on the progression of a disease.

Real Estate: Regression is employed by real estate analysts to ascertain the value of properties by considering factors such as location, size, and amenities.

Regression analysis is a critical instrument in a variety of industries and disciplines, as it offers a robust framework for making informed decisions based on data.

5.9 DIFFERENCE BETWEEN CORRELATION AND REGRESSION:

The concept of interchangeable factors is exemplified in regression analysis, whereby the relationship between the independent variable, denoted as x, and the dependent variable, denoted as y, is examined. It is important to note that the outcomes of the regression analysis may vary when the roles of x and y are interchanged. In the context of correlation, the variables x and y exhibit a property known as interchangeability, whereby they may be exchanged without altering the resulting outcome.

The distinction between correlation and regression lies in their respective representations of data. Correlation is a singular

statistical measure, often denoted as a data point, which quantifies the strength and direction of the relationship between two variables. On the other hand, regression encompasses the whole of the equation that incorporates all data points, represented by a line, to predict or model the relationship between variables.

The concept of correlation elucidates the association between two variables, while regression analysis enables us to discern the impact of one variable on another.

The relationship between variables shown by regression analysis shows a causal link. When an individual undergoes a shift, the corresponding entity also experiences a transformation, which may not always occur in a consistent manner. Correlation refers to the phenomenon in which variables have a tendency to co-vary.

5.10 LET US SUM UP

The use of simple linear regression enables researchers to estimate the parameters, namely the intercept and slopes, of linear equations that establish connections between two or more variables. Understanding the functional relationship between a dependent variable and one or more independent or explanatory factors, together with an estimation of the parameters of that relationship, significantly enhances a researcher's capacity to forecast the values that the dependent variable will assume across various circumstances. The ability to assess the impact of a single independent variable on the dependent variable, while holding other independent variables constant, may provide valuable insights for decision-making and policy formulation. The ability to empirically examine the presence of distinct impacts stemming from several independent factors is of great value to decision-makers, researchers, and policy-makers, as it enables them to discern the

relative significance of different variables. Regression analysis is a very influential statistical technique that has several advantageous properties.

The concept behind regression is straightforward: it entails determining the equation of a line that closely approximates a maximum number of data points. The mathematical principles behind regression analysis are not devoid of complexity. Rather of engaging in the process of acquiring mathematical knowledge, many researchers choose to use computers for the purpose of deriving regression equations. Consequently, the emphasis of this chapter is in the interpretation of computer-generated printouts, rather than delving into the mathematical intricacies of regression.

5.11 KEY WORDS:

Regression analysis: is a robust statistical technique that enables the examination of the association between two or more variables of interest.

Simple linear regression: enables researchers to estimate the parameters, namely the intercept and slopes, of linear equations that establish connections between two or more variables.

Curvilinear regression: refers to a regression model that aims to match a curve rather than a linear relationship.

5.12 ANSWERS TO CHECK YOUR PROGRESS

1. In regression analysis, the _____ variable is the one being predicted.
2. The dependent variable is utilized to create predictions in……...
3. The line of greatest fit in linear regression is also known as the……...

4. The ……..of determination quantifies the extent to which the independent variable(s) can account for the variability in the dependent variable.

5. The term "……." in regression analysis refers to the magnitude and direction of the association between two variables.

Answer:

1. Dependent
2. Regression analysis
3. Regression line
4. Coefficient
5. Correlation

5.13 TERMINAL QUESTIONS

Q.1. What do you mean by regression? Explain types of regression?

Q2. What are the utility of regression?

Q3. Differentiate between regression and correlation.

Q4.What do you mean by Regression coefficient? What are the Properties of Regression Coefficient?

UNIT 6: VARIOUS METHODS OF CALCULATION OF THE COEFFICIENT OF CORRELATION AND THEIR ANALYSIS (TWO VARIABLE)

Structure

6.0 Objectives

6.1 (A). Graphical Methods of Determining Correlation:

6.1.1 Scatter Diagram

6.1.2 Correlation Graph

6.2. (B). Mathematical Methods of Determining Correlation:

6.2.1. Karl Pearson's Coefficient of Correlation

6.2.2. Spearman's Rank Coefficient of correlation

6.2.3. Concurrent Deviation Method

6.3 Let Us Sum Up

6.4 Key Words

6.5 Answers to Check Your Progress

6.6 Terminal Questions

6.0 OBJECTIVES

After studying this unit, you should be able to:

- Understand various methods of calculation of Coefficient of correlation.
- Understand the concepts of Graphical Methods of Determining correlation.
- Explain Mathematical Methods of Determining correlation.
- Analyze and interpret the results of correlation analysis.

6.1 (A). GRAPHICAL METHODS OF DETERMINING CORRELATION:

6.1.1 i. Scatter Diagram

6.1.2 ii. Correlation Graph

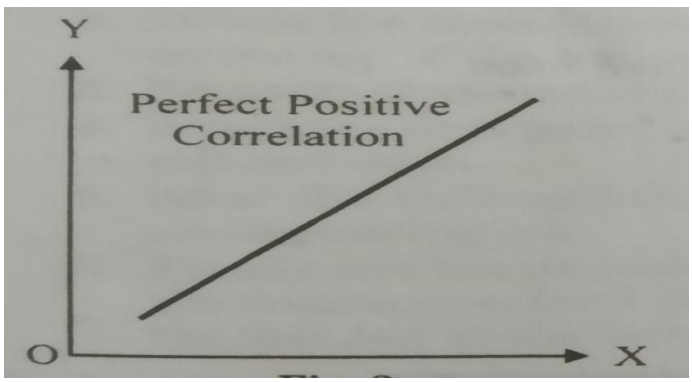
6.1.1 i. Scatter Diagram

A scatter diagram is a correlation chart that visually illustrates the relationship between two variables. The point in the graph will descend along a curve or line if the variables are correlated. A scatter diagram or scatter plot provides insight into the nature of the relationship. This diagram enables a visual assessment of the degree to which the variables are interconnected.

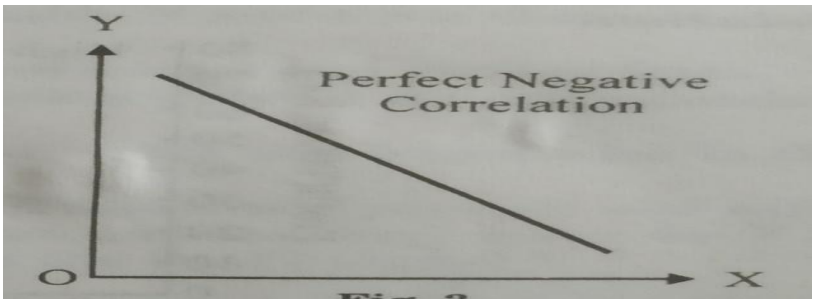
A scatter diagram is generated by measuring X on the horizontal axis and Y on the vertical axis, and then plotting a point for each pair of observations of X and Y. The diagram is composed of dots or elements.

Case 1: A linear correlation exists when the elements resemble a line.

I. The two variables exhibit a flawless positive correlation if the line ascends from the left bottom of the graph toward the right.

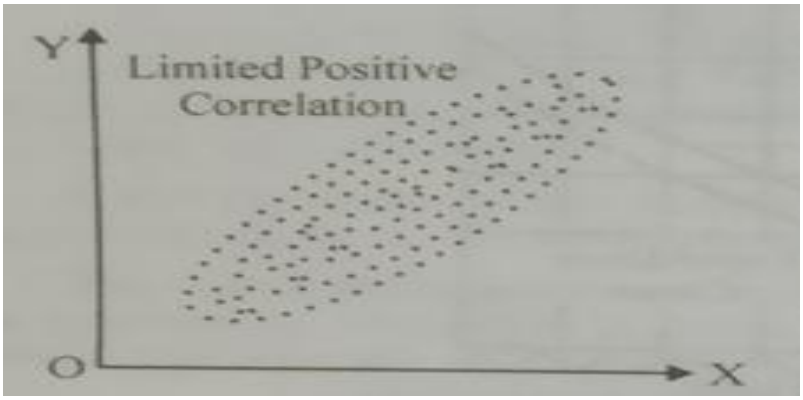


II. A perfect negative correlation between the two variables is present if the line travels in the opposite direction.

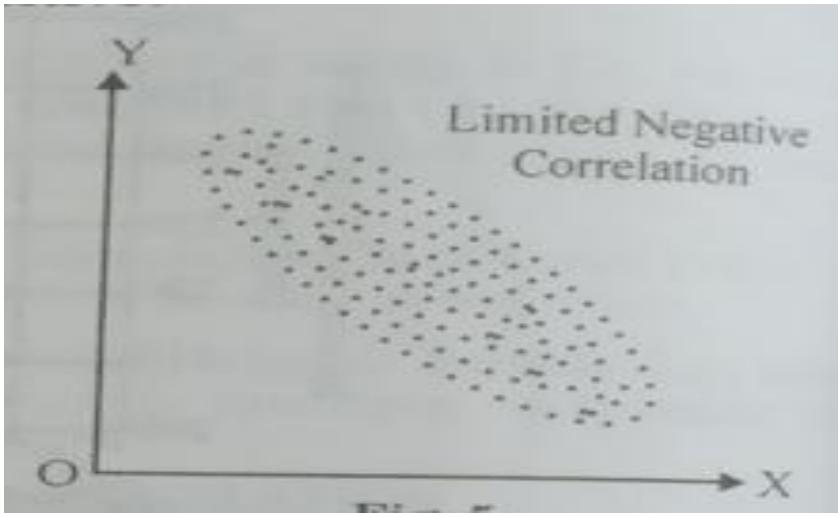


Case 2: The two variables are correlated if the points exhibit a trend, whether it be upward or downward.

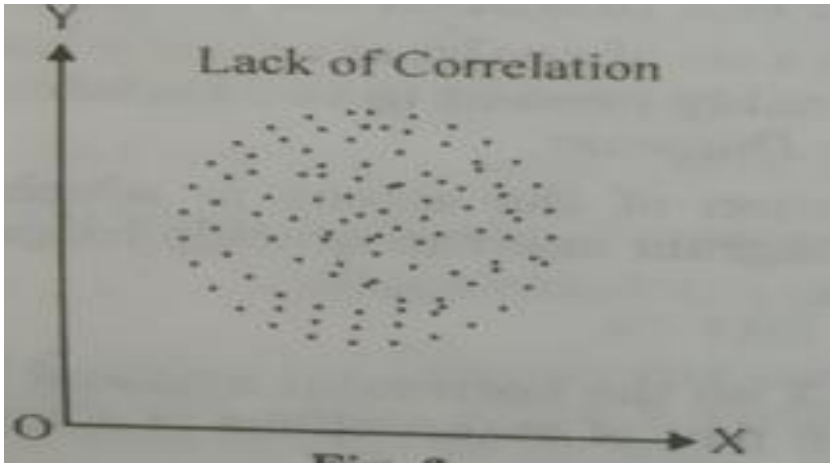
I. The correlation is positive if the figures are trending upward, ascending from the left bottom and moving upward to the right top.



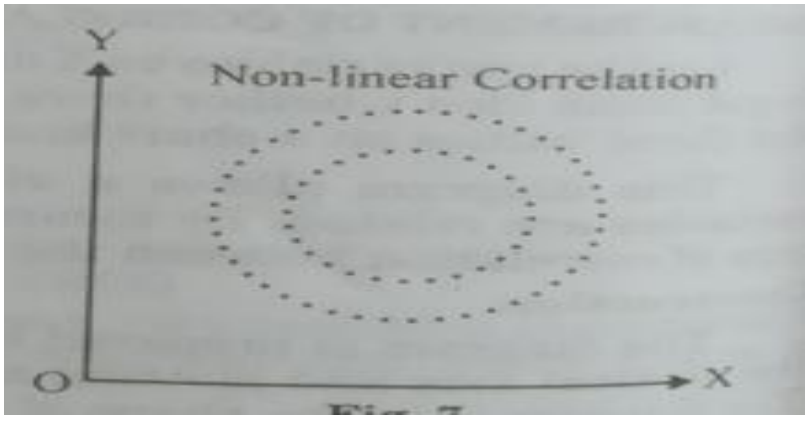
II. The correlation is negative if the tendency is reversible, resulting in a downward trend from the left to the right bottom.



Case 3: The two variables do not exhibit a correlation if the plotted points do not exhibit any trend.



Case 4: The correlation is non-linear if the points have a propensity to cluster around a circle or any other known shape other than a straight line.



6.1.2. Correlation graph

A correlation graph is a method for determining the correlation between two variables. The precise degree of correlation is not known, and this graph provides a visual study of correlation.

Two series, X and Y, are plotted on a graph paper in a correlation graph. The X-axis is used to represent the common characteristics/attributes of the two series, while the Y-axis is used to represent the two series in accordance with the appropriate scale. Different sorts of lines are employed to illustrate the two contours.

The correlation is positive when two curves are parallel and ascend from the left bottom to the right top.

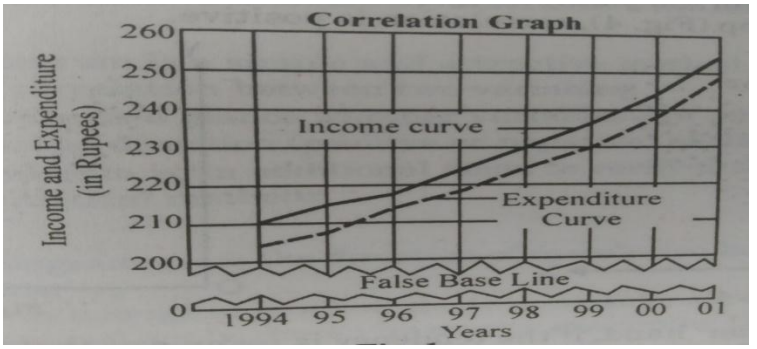
However, if they are in opposition, there is a negative correlation between them.

If the curves exhibit erratic fluctuations that indicate no similarity, they may indicate a low degree of correlation or the absence of correlation.

Example:

Year	**1994**	**1995**	**1996**	**1997**	**1998**	**1999**	**2000**	**2001**
Average income(Rs.)	**210**	**215**	**218**	**222**	**230**	**236**	**245**	**255**
Average expenditure(Rs.)	**205**	**208**	**212**	**218**	**225**	**230**	**237**	**250**

Solution:



This graph shows a very high degree of positive correlation between income and expenditure.

6.2. (B). MATHEMATICAL METHODS OF DETERMINING CORRELATION:

6.2.1. Karl Pearson's Coefficient of correlation

6.2.2. Spearmans's Rank Coefficient of correlation

6.2.3. Concurrent Deviation Method

6.2.1 Karl Pearson's Coefficient of correlation

The coefficient of correlation is determined by dividing the sum of the products of the deviations from the respective means by the number of pairs and their standard deviations, as per Karl Pearson.

It is denoted by the symbol "r," which is a pure number, meaning that it has no unit.

r = Sum of products of deviations from their respective means/Number of pairs x Their standard deviations

or

$$\mathbf{r = \frac{Sum\ of\ products\ of\ deviations\ from\ their\ respective\ means}{Number\ of\ pairs\ x\ Their\ standard\ deviations}}$$

Karl Pearson Coefficient Correlation Assumptions

A linear relationship exists between any two variables.

The outliers must be either completely eliminated or kept to a minimum range.

The Karl Pearson Coefficient of correlation's primary characteristics include:

The Correlation Coefficient (r) is not quantified.

A positive value for r indicates that both X and Y are moving in the same direction.

X and Y are traveling in contrary directions if r has a negative value.

If the value of r is 0, it is considered that X and Y are not correlated.

A large value of r suggests a robust linear relationship between two variables.

A feeble relationship between two variables is indicated by a low value of r.

A correlation between two variables is considered ideal if the value of r is either +1 or -1.

When is it appropriate to employ the Pearson correlation coefficient?

- Both variables are quantitative.
- The variables exhibit a normal distribution.
- There are no outliers in the data.
- The relationship is linear.

Karl Pearson's coefficient of correlation can be calculated using various methods.

- Actual Mean Method
- Direct Method
- Short-Cut Method /Assumed Mean Method/Indirect Method
- Step-Deviation Method

Actual Mean Method

- Calculate the mean of the given two series (say X and Y).
- Take the deviation of X series from $\bar{x}$ and denote the deviations by ($d_{\bar{x}}$).
- Square the deviations of x and obtain the total ($\sum d_{\bar{x}}2$).
- Take the deviation of Y series from $\bar{y}$ and denote the deviations by ($d_{\bar{y}}$).
- Square the deviations of y and obtain the total($\sum d_{\bar{y}}2$).
- Multiply the respective deviations of Series X and Y and obtain the total ($\sum d_{\bar{x}} d_{\bar{y}}$) .
- Apply formula to determine the Coefficient of Correlation:

$$r = \frac{\sum d_{\bar{x}} d_{\bar{y}}}{\sqrt{\sum d_{\bar{x}}2 \times \sum d_{\bar{y}}2}}$$

Direct Method

- Calculate the sum of Series X; ($\sum X$) and sum of Series Y; ($\sum Y$).
- Square the values of X Series and calculate their total; ($\sum X^2$).
- Square the values of Y Series and calculate their total; ($\sum Y^2$).
- Multiply the values of Series X and Y and calculate their total; ($\sum XY$).
- Apply formula to determine Coefficient of Correlation:

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

Short-Cut Method /Assumed Mean Method of Calculating Karl Pearson's Coefficient of Correlation:

- Take the deviations of X Series from the assumed mean and denote the values by dx. Calculate their total; ($\sum dx$).
- Square the deviations of X series and calculate their total; ($\sum dx^2$).
- Take the deviations of Y Series from the assumed mean and denote the values by dy. Calculate their total; ($\sum dy$).
- Square the deviations of Y series and calculate their total; ($\sum dy^2$).
- Multiply dx and dy and calculate their total; ($\sum dxdy$).
- Apply the following formula to determine Coefficient of Correlation:
- $$r = \frac{N\sum dxdy - \sum dx.\sum dy}{\sqrt{N\sum dx^2 - (\sum dx)^2}\sqrt{N\sum dy^2 - (\sum dy)^2}}$$ or $$r = \frac{\sum dxdy - \frac{\sum dx \times \sum dy}{N}}{\sqrt{\left[\sum d\bar{x} - \frac{(\sum dx)^2}{N}\right]\left[\sum d\bar{y} - \frac{(\sum dy)^2}{N}\right]}}$$
- N = Number of pair of observations , $\sum dx$ = Sum of deviations of X values from assumed mean, $\sum dy$ = Sum of deviations of Y values from assumed mean, $\sum dx^2$ = Sum of squared deviations of X values from assumed mean, $\sum dy^2$ = Sum of squared deviations of Y values from assumed mean, $\sum dxdy$ = Sum of the products of deviations dx and dy

Step-Deviation Method Calculating Karl Pearson's Coefficient of Correlation:

- **T**ake the deviations of Series X from the assumed mean and divide them by Common Factor (C) to determine step deviation (dx'). Calculate the total of step deviations; $\sum dx'$
- Take the deviations of Series Y from the assumed mean and divide them by Common Factor (C) to determine step deviation (dy'). Calculate the total of step deviations; $\sum dy'$
- Square the step deviation of Series X and determine their total; $\sum dx'^2$
- Square the step deviation of Series Y and determine their total; $\sum dy'^2$
- Multiply (dx') and (dy'), and determine their total; $\sum dx'dy'$
- **Apply formula:** $r = \frac{N\sum dx'dy' - \sum dx' . \sum dy'}{\sqrt{N\sum dx'^2 - (\sum dx')^2}\sqrt{N\sum dy'^2 - (\sum dy')^2}}$

(dx') =Sum of deviations of X values from assumed mean, (dy') = Sum of deviations of Y values from assumed mean,

$\sum dx'^2$ =Sum of squared deviations of X values from assumed mean,

$\sum dy'^2$ = Sum of squared deviations of Y values from assumed mean,

$\sum dx'dy'$ = Sum of the products of deviations $\sum dx'$ and $\sum dy'$.

Example. 1.

Calculate Coefficient of Correlation from the following data:

X	50	16	36	46	30	40	20	26
Y	80	35	57	76	60	66	40	50

Solution:

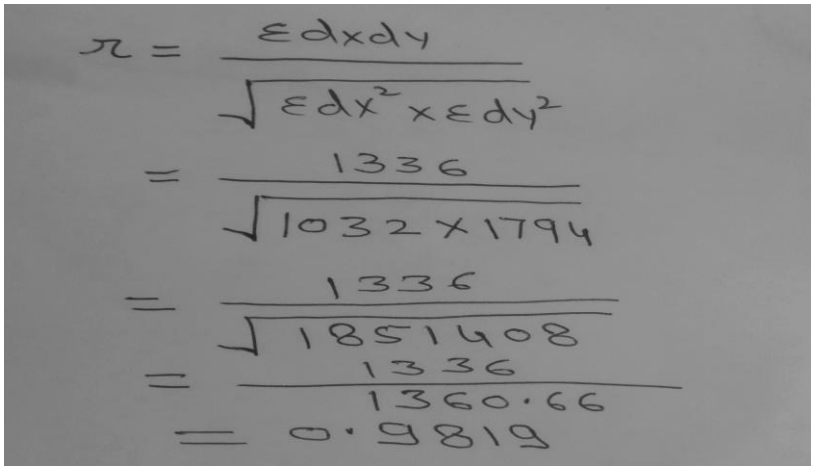
$\bar{X} = \frac{264}{8}$=33, $\bar{Y} = \frac{464}{8}$=58

X	**dx(33)**	**dx2**	**Y**	**dy(58)**	**dy2**	**dxdy**
50	17	289	80	22	484	374
16	-17	289	35	-23	529	391
36	3	9	57	-1	1	-3
46	13	169	76	18	324	234
30	-3	9	60	2	4	-6

40	7	49	66	8	64	56
20	-13	169	40	-18	324	234
26	-7	49	50	-8	64	56
∑X=264		**∑dx2=1032**	**∑X=464**		**∑dy2=1794**	**∑dxdy = 1336**

Apply Formula:

$$\mathbf{r} = \frac{\sum d_{\bar{x}} d_{\bar{y}}}{\sqrt{\sum d_{\bar{x}}2 \ \mathbf{x} \sum d_{\bar{y}}2}}$$



Value of Coefficient of correlation = 0.9819

Example. 2.

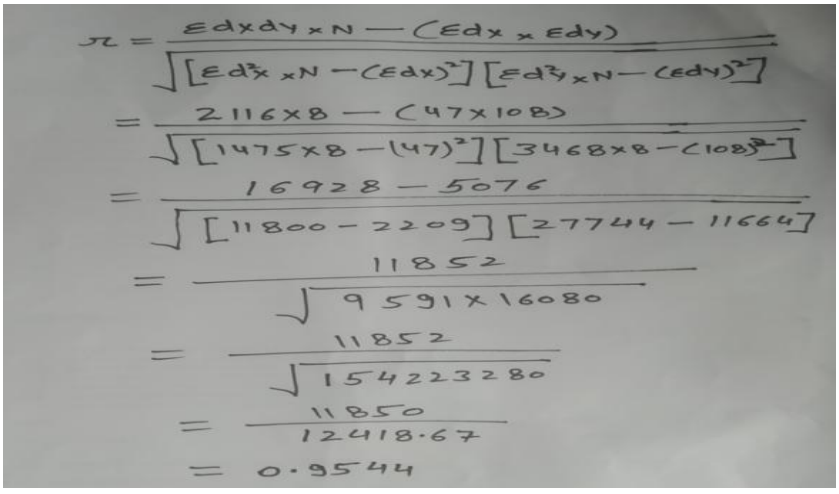
Calculate Coefficient of Correlation from the following data:

X	78	89	96	69	59	79	68	61
Y	125	137	156	112	107	136	123	108

Solution:

X	**dx(69)**	**dx2**	**Y**	**dy(112)**	**dy2**	**dxdy**
78	9	81	125	13	169	117
89	20	400	137	25	625	500

96	27	729	156	44	1936	1188
69	0	0	112	0	0	0
59	-10	100	107	-5	25	50
79	10	100	136	24	576	240
68	-1	1	123	11	121	-11
61	-8	64	108	-4	16	32
	∑dx = 47	**∑dx2 = 1475**		**∑ dy = 108**	**∑dy2 =3468**	**∑dxdy = 2116**



Value of Coefficient of Correlation = 0.9544

Example. 3.

X	44	46	46	48	52	54	54	_	60	60
Y	36	40	42	40	_	44	46	48	50	52

Calculate Coefficient of Correlation, taking deviations from actual mean $\bar{X}$ =52 and $\bar{Y}$ = 44 of the following data:

Solution:

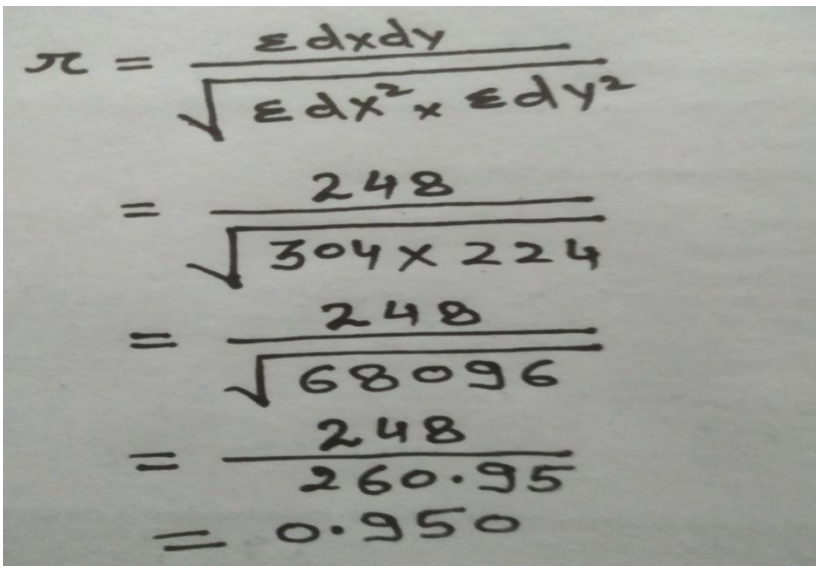
Actual mean are given in this question. So missing values can be calculated on the basis of sum of the values in series X and Y will

be ($\sum X = 52x\ 10 - 520$) and ($\sum Y = 44x\ 10 = 440$). On the basis of total of series missing value:

Series X = 520 - (44+46+46+48+52+54+56+60+60) = 520 - 466 = 54

Series Y = 440 – (36+40+42+40+44+46+48+50+52) = 440 – 398 = 42

X	dx(52)	dx2	Y	dy(44)	dy2	dxdy
44	-8	64	36	-8	64	64
46	-6	36	40	-4	16	24
46	-6	36	42	-2	4	12
48	-4	16	40	-4	16	16
52	0	0	42	-2	4	0
54	2	4	44	0	0	0
54	2	4	46	2	4	4
56	4	16	48	4	16	16
60	8	64	50	6	36	48
60	8	64	52	8	64	64
	∑dx = 0	**∑dx2 = 304**		**∑ dy = 0**	**∑dy2 =224**	**∑dxdy = 248**



Value of Coefficient of Correlation = 0.950

6.2.2. Spearman's rank coefficient of correlation

- The Spearman's rank coefficient of correlation or Spearman correlation coefficient is a nonparametric measure of rank correlation.
- Spearman suggested ranking the values of X and also ranking the values of Y. These ranks are then used instead of the actual values of X and Y in the formula for Spearman's rank correlation coefficient. The result of this calculation is the Spearman rank correlation coefficient. It is often denoted by the Greek letter 'ρ' (rho) and is primarily used for data analysis.

Formula :

$$\rho = 1 - \frac{6\sum D^2}{N(N^2-1)}$$

- ρ = Spearman Correlation coefficient
- $\sum D^2$ = Squares of the difference in the rank of each pair of observations.
- **N** = Total number of observation

Calculation of Spearman's rank Coefficient of Correlation

When ranks of items are given in the question:-

- Calculate $(R_x - R_y)$ and it is shown under the column headed by D.
- These differences are squared up (D^2) and $(\sum D^2)$ is obtained by totalling these squares and finally apply formula.

Example. 1.

Calculate Rank Coefficient of Correlation from the following data:

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

- **Solution:**

Rank X (Rx)	Rank Y (Ry)	D =Rx - Ry	D2
5	3	2	4
3	7	-4	16
4	5	-1	1

8	**9**	**-1**	**1**
2	**2**	**0**	**0**
1	**4**	**-3**	**9**
7	**1**	**6**	**36**
10	**10**	**0**	**0**
6	**8**	**-2**	**4**
9	**6**	**3**	**9**
N =10		**∑ D = 0**	**∑D2 = 80**

- $r_r = 1 - \frac{6\sum D^2}{N(N^2-1)}$
- $r_r = 1 - \frac{6\ x\ 80}{10(100-1)}$
- $r_r = 1 - \frac{6\ x\ 80}{10(99)}$
- $r_r = 1 - \frac{4\ 80}{990}$
- $r_r = \frac{990 - 4\ 80}{990}$
- $r_r = \frac{550}{990}$
- $r_r = 0.52$

When ranks are not given:-

Ranks are assigned to the values given in both the series. These ranks may be assigned in ascending or descending order but generally descending order is preferred in which the highest value is given rank 1 and the other values are given rank accordingly. After assigning the ranks differences of Ranks of X (R_x) and ranks of Y (R_y) is calculated (R_x - R_y) and it is shown under the column headed by D. These differences are squared up (D^2) and ($\sum D^2$) is obtained by totalling these squares and finally apply formula.

$$\rho = 1 - \frac{6\sum D^2}{N(N^2-1)}$$

Example. 2.

Calculate Rank Coefficient of Correlation from the following data:

X	**20**	**22**	**24**	**25**	**30**	**32**	**28**	**21**	**26**	**35**
Y	**16**	**15**	**20**	**21**	**19**	**18**	**22**	**24**	**23**	**25**

Solution:

X	**Rank X (Rx)**	**Y**	**Rank Y (Ry)**	**D =Rx - Ry**	**D2**
20	**10**	**16**	**9**	**1**	**1**
22	**8**	**15**	**10**	**-2**	**4**
24	**7**	**20**	**6**	**1**	**1**
25	**6**	**21**	**5**	**1**	**1**
30	**3**	**19**	**7**	**-4**	**16**
32	**2**	**18**	**8**	**-6**	**36**
28	**4**	**22**	**4**	**0**	**0**
21	**9**	**24**	**2**	**7**	**49**
26	**5**	**23**	**3**	**2**	**4**
35	**1**	**25**	**1**	**0**	**0**
N =10				**∑ D = 0**	**∑D2 = 112**

$$r_r = 1 - \frac{6\sum D^2}{N(N^2-1)}$$

$$r_r = 1 - \frac{6\ x\ 112}{10(100-1)}$$

$$r_r = 1 - \frac{6\ x\ 112}{10(99)}$$

$$r_r = 1 - \frac{672}{990}$$

$$r_r = \frac{990-672}{990}$$

$$r_r = \frac{318}{990}$$

$$r_r = 0.32$$

When some values or ranks are equal:-

- i. Bracket Rank Method ii. Average Rank Method

i. Bracket Rank Method:

- This method, the same rank is assigned to all items of that same value, which is given to the first item of the value. After this value, that rank is given to the next value which would have been assigned in case of difference in the items of the same value.
- Ranks are to be assign to the values 12, 15, 11, 15, 18 and 15, then rank 1 will be given to 18, rank 2 to all the three items of 15, rank 5 to 12 and rank 6 to 11 hear the rank 5 is given to 12 because if the previous three items (15) are not equal then their ranks would also have been 2, 3 and 4 in place of 2.

$$\rho = 1 - \frac{6[\sum D^2 + \frac{1}{12}(m^3 - m) + \frac{1}{12}(m^3 - m) + \ldots\ldots\ldots]}{N(N^2 - 1)}$$

ii. Average Rank Method:

In this method average rank is assigned to the items of the same value.

- In this the values 12,15 ,11,15 ,18 and 15 rank 3 will be assigned to the value 15. The basis is that rank 1 will be assigned to 18, after that the value 15 has come thrice and if the separate rank would have been assigned they would have been 2, 3 and 4 but due to equal value equal rank is to be assigned and that will be the average of 2 , 3 and 4. i.e. $\frac{2+3+4}{3} = 3$, after this rank 5 and 6 will be assigned to 12 and 11 respectively.

$$\boldsymbol{\rho = 1 - \frac{6[\sum D^2 + \frac{1}{12}(m^3 - m) + \frac{1}{12}(m^3 - m) + \ldots\ldots\ldots]}{N(N^2 - 1)}}$$

ρ = Spearman Correlation coefficient, $\sum D^2$ = Squares of the difference in the rank of each pair of observations, **N** = Total number of observation, m = Number of times an item is repeated

Example. 3.

Calculate Rank Coefficient of Correlation from the following data:

X	**22**	**24**	**27**	**35**	**21**	**20**	**27**	**25**	**27**	**23**
Y	**30**	**38**	**40**	**50**	**38**	**25**	**38**	**36**	**41**	**32**

X	**Rank X (Rx)**	**Y**	**Rank Y (Ry)**	**D =Rx - Ry**	**D2**
22	**8**	**30**	**9**	**-1**	**1**
24	**6**	**38**	**5**	**1**	**1**
27	**3**	**40**	**3**	**0**	**0**
35	**1**	**50**	**1**	**0**	**0**
21	**9**	**38**	**5**	**4**	**16**
20	**10**	**25**	**10**	**0**	**0**
27	**3**	**38**	**5**	**-2**	**4**
25	**5**	**36**	**7**	**-2**	**4**
27	**3**	**41**	**2**	**1**	**1**
23	**7**	**32**	**8**	**-1**	**1**
N =10				**∑ D = 0**	**∑D2 = 28**

$\frac{2+3+4}{3} = 3$ **and** $\frac{4+5+6}{3} = 5$

$$r_r = 1 - \frac{6[\sum D^2 + \frac{1}{12}(m^3 - m) + \frac{1}{12}(m^3 - m)]}{N(N^2 - 1)}$$

$$r_r = 1 - \frac{6[28 + \frac{1}{12}(3^3 - 3) + \frac{1}{12}(3^3 - 3)]}{10(10^2 - 1)}$$

$$r_r = 1 - \frac{6[28 + \frac{1}{12}(27 - 3) + \frac{1}{12}(27 - 3)]}{10(100 - 1)}$$

$$r_r = 1 - \frac{6[28 + \frac{1}{12}(24) + \frac{1}{12}(24)]}{10(99)}$$

$$r_r = 1 - \frac{6[28 + 2 + 2]}{990}$$

$$r_r = 1 - \frac{6[28 + 4]}{990}$$

$$r_r = 1 - \frac{6 \, x \, 32}{990}$$

$$r_r = 1 - \frac{192}{990}$$

$$r_r = \frac{990 - 192}{990}$$

$$r_r = \frac{798}{990}$$

$$r_r = 0.81$$

6.2.3 Concurrent deviation method for calculation Coefficient of Correlation

Concurrent deviation method is the simplest method of studying correlation. It is generally used at that time when the main objective of study is to find out the direction of correlation because in its calculation the magnitude of deviation is ignored. Only direction of deviation (increase or decrease, positive or negative) is taken into account.

Process of calculating coefficient of correlation by concurrent deviation method:-

i. First of all, deviation signs or marked in both the series on the basis of direction of change. These signs are marked on the basis of comparison of each value with its preceding value. In this process no sign is marked for the first value because it has no proceeding value. Now the second value is compared with the first value. If the second value is increasing put a sign of Plus (+), If it is decreasing put a sign of minus (-) and if it is constant , put sign of

equal (=). Similarly the third value will be compared with the second value and this process continues for all other values. It may be noted that only sign of deviation is to be marked and the magnitude of deviation is not shown. These deviations are put in the column of dx and dy for the series of X and Y respectively.

ii. Deviation signs of both the series are multiplied and the sign of such multiplication is put in the last column. The value of 'C' concurrent is determined on the basis of positive signs of dxdy

iii. Apply Formula:

$$r_c = \pm \sqrt{\pm[\frac{2C-n}{n}]}$$

- n = N-1, N is the total number of pairs of observation in x and y series
- C = Value of concurrent

Example. 4.

Calculate Coefficient of Correlation by Concurrent deviation method from the following data:

Supply	90	95	96	94	96	100	102	104	105	108	110
Price	140	130	132	134	130	128	124	120	117	118	110

Solution:

X	**Direction of change (Dx)**	**Y**	**Direction of change (Dy)**	**C = Dx X Dy**
90		**140**		
95	+	**130**	-	
96	+	**132**	+	+

94	**-**	**134**	+	
96	+	**130**	**-**	
100	+	**128**	**-**	
102	+	**124**	**-**	
104	+	**120**	**-**	
105	+	**117**	**-**	
108	+	**118**	+	+
110	+	**110**	**-**	
	n = 10		**n = 10**	**C=2**

$$r_c = \pm\sqrt{\pm \frac{(2C-n)}{n}}$$

$$r_c = \pm\sqrt{\pm[\frac{(2x2-10)}{10}]}$$

$$r_c = \pm\sqrt{\pm[\frac{(4-10)}{10}]}$$

$$r_c = -\sqrt{-[\frac{-6}{10}]}$$

$$r_c = -\sqrt{-(-0.6)}$$

$$r_c = -\sqrt{0.6}$$

$$r_c = -0.78$$

6.3 LET US SUM UP

Pearson's correlation coefficient: The linear relationship between two continuous variables is quantified by the Pearson correlation coefficient (r). It is within the range of -1 to 1, where:

A perfect positive linear relationship is denoted by +1, a perfect negative linear relationship by -1, and no linear relationship by 0.

Spearman's rank correlation coefficient (ρ or rs) quantifies the intensity and direction of the monotonic relationship between two ranked variables. It is a non-parametric test, which implies that it does not presuppose that the data is normally distributed.

The concurrent deviation method is the most straightforward approach to correlation analysis. It is typically employed during the phase of the investigation in which the primary objective is to determine the direction of the correlation, as the magnitude of the deviation is omitted from the calculation. Only the direction of deviation (positive or negative, increase or decrease) is considered.

6.4 KEY WORDS:

Correlation coefficient: is a statistical measure that describes the strength and direction of a relationship between two variables.

Scatter diagram: is a correlation chart that visually illustrates the relationship between two variables.

6.5 ANSWERS TO CHECK YOUR PROGRESS

1. A widely used approach to quantify the magnitude and direction of a linear association between two continuous variables is the _____ correlation coefficient, typically represented as 'r'.
2. Another approach, which quantifies the intensity and direction of a monotonic association between two variables, is the _____ correlation coefficient, often represented as 'ρ'.
3. A Pearson's correlation value of 0 shows ……..connection.
4. To compute Pearson's correlation coefficient, both variables must be…….

5. A …….. Spearman's rank correlation value shows that when one variable increases, the other variable decreases.

3. Answer:

1. Pearson
2. Spearman
3. no linear
4. continuous
5. negative

6.6 TERMINAL QUESTIONS:

Q1. Calculate Rank Coefficient of Correlation from the following data.

Rank X	**1**	**5**	**2**	**3**	**4**	**8**	**7**	**6**
Rank Y	**5**	**2**	**3**	**7**	**4**	**8**	**6**	**1**

Q.2. Calculate Rank Coefficient of Correlation from the following data.

X	**10**	**17**	**9**	**10**	**31**	**25**	**21**	**12**
Y	**9**	**20**	**15**	**25**	**23**	**32**	**15**	**19**

Q.3. Calculate Coefficient of Correlation by Concurrent deviation method from the following data:

X	**43**	**23**	**32**	**24**	**26**	**27**	**29**	**28**	**20**	**40**
Y	**18**	**20**	**21**	**20**	**21**	**22**	**23**	**24**	**25**	**26**

Q4. Calculate Karl Pearson's Coefficient of Correlation from the following data.

X	**200**	**64**	**144**	**184**	**120**	**160**	**80**	**104**
Y	**320**	**140**	**228**	**304**	**240**	**264**	**160**	**200**

Q.5. Calculate Karl Pearson's Coefficient of Correlation from the following data.

X	**156**	**178**	**192**	**138**	**118**	**158**	**136**	**122**
Y	**250**	**274**	**312**	**224**	**214**	**272**	**246**	**216**

Q.6. Calculate Coefficient of Correlation, taking deviations from actual mean $\bar{X}$=26 and $\bar{Y}$= 22 of the following data:

X	**22**	**23**	_	**24**	**26**	**27**	**27**	**28**	**20**	**40**
Y	**18**	**20**	**21**	**20**	**21**	**22**	**23**	**24**	_	**26**

UNIT 7: REGRESSION ANALYSIS

Structure

7.0 Introduction

7.1 Objectives

7.2 Regression lines

7.3 Justifications for the existence of two regression lines:

7.4 Functions of regression lines

7.5 Regression equations

7.6 Example of Regression equation:

7.7 Regression Coefficient

7.8 Properties of Regression Coefficient

7.9 Let Us Sum Up

7.10 Key Words

7.11 Answers to Check Your Progress

7.12 Terminal Questions

7.0 INTRODUCTION:

The purpose of regression analysis is to establish a relationship between related variables in order to estimate or predict the corresponding value of the other variable based on the value of the first variable.

M.M. Blair defines regression analysis as "A mathematical measure of the average relationship between two or more variables in terms of the original unit of data."

As per Wallies and Roberts, "It is frequently more crucial to ascertain the actual nature of the relationship in order to estimate or predict one variable (the dependent variable). The statistical technique that is most suitable for this situation is regression analysis."

7.1 OBJECTIVES

After studying this unit, you should be able to:

- Understand concept of Regression lines.
- Describe Functions of regression lines.
- Learn how to develop Regression equations.

Explain Properties of Regression Coefficient

7.2 REGRESSION LINES

The lines of greatest fit that convey the mutual average relationship between two series are known as regression lines. These lines provide the most accurate estimate of one variable for any given value of the other variable.

In the context of regression analysis, having two regression lines typically refers to the regression line of y on x and x on y .

Based on x, the regression line of y on x forecasts the value of y.

Based on y, the regression line of x on y forecasts the value of x.

7.3 JUSTIFICATIONS FOR THE EXISTENCE OF TWO REGRESSION LINES:

1. Modify the independent and dependent variables

2. Minimization of Errors

3. A more comprehensive comprehension of the relationship between the two variables is facilitated by the presence of two regression lines, which offer insight into the nature of the dependency in both directions.

4. Each regression line accounts for the error in one variable while presuming that the other variable is measured without error, provided that there is measurement error in both variables.

7.4 FUNCTIONS OF REGRESSION LINES:

The regression lines execute the subsequent function:

1. To denote the direction and magnitude of the correlation

2. In order to obtain the most accurate estimate

3. Calculation of the mean value

4. Variance ratio

7.5 REGRESSION EQUATIONS

Regression equations are algebraic expressions of the regression lines, and since there are two regression lines, there are two regression equations.
1. Regression equation of X on Y: This equation is employed to estimate the value of X for the given value of Y.
2. Equation of regression for Y with respect to X
This equation is employed to determine the value of Y in relation to the given value of X.

- **1. Regression equation of X on Y:**
- This equation is used for estimating the value of x for the given value of Y.

- $(X-\bar{X}) = r\frac{\sigma X}{\sigma Y}(Y-\bar{Y})$
- X=Value of x variable to be predicted,
- $\bar{X}$ = Arithmetic mean of X series,
- $\bar{Y}$ = Arithmetic mean of by Y series
- r = correlation coefficient of x and y series,
- σ_X =Standard deviation of x series,
- σ_Y =standard deviation of y series
- The value of Y variable corresponding to which the value of x variable is to be predicted.
- **2. Regression equation of Y on X:**
- This equation is used for estimating the value of Y for the given value of X.
- $(Y-\bar{Y}) = r\frac{\sigma Y}{\sigma X}(X-\bar{X})$
- X=Value of x variable to be predicted, $\bar{X}$ = Arithmetic mean of X series, $\bar{Y}$ = Arithmetic mean of by Y series
- r = correlation coefficient of x and y series, σ_X =Standard deviation of x series, σ_Y =standard deviation of y series
- The value of X variable corresponding to which the value of Y variable is to be predicted

7.6 EXAMPLE OF REGRESSION EQUATION:

Example:

Obtain both regression equations from the following information:

	X-series	Y-series
Mean	**18**	**100**
S.D.	**14**	**20**

Coefficient of Correlation Between X and Y series (r) = + 0.8

Solution:

Regression equation X on Y

- **(X-$\bar{X}$)= $r\frac{\sigma X}{\sigma Y}$ (Y-$\bar{Y}$)**
- $\bar{X}$ = Arithmetic mean of X series= 18
- $\bar{Y}$ = Arithmetic mean of by Y series = 100
- r = correlation coefficient of x and y series= 0.8
- σ_X =Standard deviation of x series= 14
- σ_Y =standard deviation of y series= 20
- **(X-$\bar{X}$)= $r\frac{\sigma X}{\sigma Y}$ (Y-$\bar{Y}$)**
- **(X-18) = $0.8\frac{14}{20}$(Y-100)**
- **(X-18) =0.56 (Y-100)**
- **(X-18) = 0.56Y - 56**
- **X = 0.56Y - 56 + 18**
- **X = 0.56Y - 38**
- **Regression equation Y on X**
- $(Y-\bar{Y}) = r\frac{\sigma Y}{\sigma X}$ (X-$\bar{X}$)
- $\bar{X}$ = Arithmetic mean of X series= 18
- $\bar{Y}$ = Arithmetic mean of by Y series = 100
- r = correlation coefficient of x and y series= 0.8
- σ_X =Standard deviation of x series= 14
- σ_Y =standard deviation of y series= 20
- $(Y-\bar{Y}) = r\frac{\sigma Y}{\sigma X}$ (X-$\bar{X}$)
- **(Y-100)= $0.8\frac{20}{14}$ (X-18)**
- **(Y-100)= 1.143 (X-18)**
- **(Y-100)= 1.143 X – 20.574**
- **Y= 1.143 X – 20.574 + 100**
- **Y= 1.143 X + 79.426**

Example:

The following information are given to you:

	X- SERIES		Y- SERIES
MEAN	**20**		**100**
S.D.	**15**		**20**
Coefficient of correlation (r)	**0.8**		

Find the most probable value of Y if x is 30 and the most probable value of x if Y is 90.

Solution:

Regression equation X on Y

- This equation is used for estimating the value of x for the given value of Y.
- $(X-\bar{X}) = r\frac{\sigma X}{\sigma Y}(Y-\bar{Y})$
- X=Value of x variable to be predicted
- Y = 90
- $\bar{X}$ = Arithmetic mean of X series= 20
- $\bar{Y}$ = Arithmetic mean of by Y series = 100
- r = correlation coefficient of x and y series= 0.8
- σ_X =Standard deviation of x series= 15
- σ_Y =standard deviation of y series= 20
- This equation is used for estimating the value of X for the given value of Y.
- $(X-20) = 0.8\frac{15}{20}(Y-100)$
- (X-20) =0.6 (Y-100)
- (X-20) =0.6 (90-100)
- (X-20) =0.6 (-10)
- (X-20) = -6
- X= 20 - 6
- X = 14
- **Regression equation Y on X**

This equation is used for estimating the value of Y for the given value of X.

- $(Y - \bar{Y}) = r\frac{\sigma Y}{\sigma X}$ (X-$\bar{X}$)
- Y =Value of Y variable to be predicted
- X = 30
- $\bar{X}$ = Arithmetic mean of X series= 20
- $\bar{Y}$ = Arithmetic mean of by Y series = 100
- r = correlation coefficient of x and y series= 0.8
- $\boldsymbol{\sigma}_X$ =Standard deviation of x series= 15
- $\boldsymbol{\sigma}_Y$ =standard deviation of y series= 20
- This equation is used for estimating the value of Y for the given value of X.
- (Y-100)= $0.8\frac{20}{15}$ (X-20)
- (Y-100)= 1.067 (X-20)
- (Y-100)= 1.067 (30-20)
- (Y-100)= 1.067 (10)
- (Y-100)= 10.67
- Y= 100 + 10.67
- Y = 110.67

Example:

Obtain both regression equations from the following information:

X	6	2	10	4	8
Y	9	11	5	8	7

Solution:

X	dx from 6	dx2	Y	dy from 8	dy2	dxdy
6	0	0	9	1	1	0
2	-4	16	11	3	9	-12
10	4	16	5	-3	9	-12

4	-2	4	8	0	0	0
8	2	4	7	-1	1	-2
∑X= 30		∑dx2= 40	∑Y= 40		∑dy2= 20	∑dxdy = -26

- Regression equation X on Y
- $\mathbf{(X-\bar{X})= \frac{\sum dxdy}{\sum dy^2} (Y-\bar{Y})}$
- $\bar{X}$ = Arithmetic mean of X series= 6
- $\bar{Y}$ = Arithmetic mean of by Y series = 8
- ∑dxdy = -26
- $\sum dy^2 = 20$
- $\mathbf{(X-\bar{X})= \frac{\sum dxdy}{\sum dy^2} (Y-\bar{Y})}$
- $\mathbf{(X-6) = \frac{-26}{20} (Y- 8)}$
- $\mathbf{(X-6) = -1.3 (Y- 8)}$
- $\mathbf{(X- 6) = -1.3Y + 10.4}$
- $\mathbf{X = -1.3Y + 10.4 + 6}$
- $\mathbf{X = -1.3Y + 16.4}$

Regression equation Y on X

- $\mathbf{(Y - \bar{Y}) = \frac{\sum dxdy}{\sum dx^2} (X-\bar{X})}$
- $\bar{X}$ = Arithmetic mean of X series= 6
- $\bar{Y}$ = Arithmetic mean of by Y series = 8
- ∑dxdy = -26
- $\sum dx^2 = 40$
- **Regression equation Y on X**
- $\mathbf{(Y - \bar{Y}) = \frac{\sum dxdy}{\sum dx^2} (X-\bar{X})}$
- $\mathbf{(Y - 8) = \frac{-26}{40} (X- 6)}$
- $\mathbf{(Y- 8) = -0.65 (X - 6)}$
- $\mathbf{(Y- 8) = -0.65 X + 3.9}$
- $\mathbf{Y = -0.65 X + 3.9 + 8}$

- $\mathbf{Y = -0.65\,X + 11.9}$

7.7 REGRESSION COEFFICIENT

The regression coefficient is an algebraic measurement of the slope of the regression line. This coefficient denotes that the average change in the values of the other variables will be determined by a unit change in the value of one variable. Given that there are two regression equations, there are also two regression coefficients.

- 1. Regression coefficient of X on Y (b_{XY}) :
- 2. Regression coefficient of Y on X (b_{YX}) :

1. Regression coefficient of X on Y (b_{XY}) :

This Coefficient represents the change in the value of variable X for a unit change in the value of the variable Y.

- $(b_{XY}) = r\frac{\sigma X}{\sigma Y}$

2. Regression coefficient of Y on X (b_{YX}) :

This Coefficient represents the change in the value of variable Y for a unit change in the value of the variable X.

- $(b_{YX}) = r\frac{\sigma Y}{\sigma X}$

7.8 PROPERTIES OF REGRESSION COEFFICIENT

1. Same sign
2. Value of coefficient
3. Calculation of correlation coefficient
4. Sign of correlation coefficient
5. Mean and coefficient of correlation
6. Regression Coefficient are independent of change of origin but not on scale.

7.9 LET US SUM UP

Regression analysis is a statistical method used to examine and construct a mathematical representation of the connection between a dependent variable and one or more independent variables. It is often used for prediction and forecasting, as well as for assessing the magnitude and characteristics of correlations between variables.

Regression analysis encompasses several types of statistical models used to analyze the relationship between a dependent variable and one or more independent variables. These models include linear regression, multiple regression, polynomial regression, logistic regression, and time series regression, among others.

1. Simple Linear Regression: Utilizes just one independent variable.
2. Multiple Linear Regression: Includes two or more independent variables.

Procedure for Regression Analysis

1. Specify the Problem: Determine the variables that are dependent and independent.
2. Data Collection: Collect data that is relevant to the variables.
3. Data Preparation: Clean and preprocess the data by handling missing values and encoding category variables if needed.
4. Model Selection: Select the suitable regression model type depending on the data and the situation at hand.
5. Model Fitting: Utilize statistical tools or programming languages such as R or Python to accurately fit the model to the given data.
6. Model Evaluation: Evaluate the model's performance by using measures such as R-squared, Adjusted R-squared, Mean Squared Error (MSE), etc.

7. Interpretation: Examine the coefficients and ascertain the importance and influence of each independent variable.
8. Forecasting: Utilize the model to generate forecasts on novel data.

7.10 KEY WORDS

Coefficients: Represent the amount by which the dependent variable changes when the independent variable increases by one unit.

Intercept: The value of the dependent variable when all independent variables have a value of zero.

p-value: is a statistical measure that assesses the likelihood of the null hypothesis being true, which states that a coefficient has no impact and is equal to zero.

7.11 ANSWERS TO CHECK YOUR PROGRESS:

1. In regression analysis, the term "……..." refers to the assumption that the errors have a constant variance across all levels of the independent variable.
2. If the regression line represents the relationship between X and Y, then X is referred to as the ……variable.
3.If bxy and byx are two regression coefficients, they possess……. Signs.
4. If one regression coefficient is negative, the other is also…….
5. The regression coefficient and correlation coefficient of the two variables will be same if their values are ……..

Answer:

1. Homoscedasticity
2. Independent

3. Same
4. Negative
5. Identical.

7.12 TERMINAL QUESTIONS:

Q.1. What do you mean by regression? Explain types of regression?

Q2. What are the utility of regression?

Q3. Differentiate between regression and correlation.

Q4.What do you mean by Regression coefficient? What are the Properties of Regression Coefficient?

Q5. Obtain both regression equations from the following information:

X	16	12	20	14	18
Y	19	22	10	18	14

Q6. The following information are given to you:

	X- SERIES		**Y-SERIES**
MEAN	**40**		**60**
S.D.	**15**		**10**
Coefficient of correlation (r)	**0.9**		

Find the most probable value of Y if x is 20 and the most probable value of x if Y is 30.

BLOCK III: ANALYSIS OF TIME SERIES

UNIT 8: CONCEPT OF TIME SERIES ANALYSIS, ADDITIVE AND MULTIPLICATION MODEL

Structure

8.0 Introduction

8.1 Objectives

8.2 Why we use Time series analysis?

8.3 Components of Time Series Data

8.4 Important Time Series Terms & Concepts

8.5 Time Series Analysis Techniques

8.6 Advantages of time series analysis

8.7 Disadvantages of time series analysis

8.8 Additive model

8. 9 Multiplicative model

8.10Difference between Additive model and Multiplicative model

8.11 Let Us Sum Up

8.12 Key Words

8.13 Answers to Check Your Progress

8.14 Terminal Questions

8.0 INTRODUCTION

The study of time series is a robust statistical technique that investigates data points gathered at consistent intervals to reveal hidden patterns and trends. This approach is widely applicable in many sectors, since it facilitates informed decision-making and

precise forecasting using historical data. Time series analysis is of utmost importance in several domains like finance, healthcare, energy, management of supply chains, weather prediction, marketing, and more, since it involves comprehending historical data and making future predictions. This tutorial will provide an in-depth exploration of time series analysis, including its definition, purpose, significance, structure, and fundamental ideas. It aims to provide you with the necessary knowledge to effectively use time series in your data analytics endeavours.

Analysis of time series is essential in the fields of data science, statistics, and analytics.

Time series analysis is primarily concerned with the examination and interpretation of a succession of data points that have been recorded or gathered at regular time intervals. Time series data differs from cross-sectional data in that it represents a dynamic and developing sequence of events throughout time, ranging from short to very long periods. An study of this kind is crucial for revealing the fundamental structures present in the data, including patterns, recurring patterns, and fluctuations that occur throughout certain seasons.

Time series analysis is to mathematically represent the underlying patterns in the data, taking into consideration factors such as autocorrelation, seasonal variations, and trends. The sequential arrangement of data points is of utmost importance; altering their order may result in the loss of significant insights or the distortion of conclusions. In addition, time series analysis often need a large dataset in order to maintain the statistical significance of the results. Analysts may use this feature to exclude irrelevant data, ensuring that the observed patterns are not random occurrences but rather statistically meaningful trends or cycles.

In order to have a more comprehensive understanding of the topic, it is necessary to differentiate between time-series data, time-series forecasting, and time-series analysis. Time-series data is a collection of observations arranged in chronological order. Conversely, time-series forecasting utilizes past data to generate future predictions, often applying statistical models such as ARIMA (Auto Regressive Integrated Moving Average). Time series analysis is a comprehensive approach that examines data in order to detect and predict its inherent patterns, such as seasonality, trends, and cycles. Time series distinguishes itself by its temporal dependence, the need for a suitably extensive dataset for precise analysis, and its distinctive ability to emphasize evolving cause-effect linkages.

8.1 OBJECTIVES

After studying this unit, you should be able to:

- Explain concept of Time Series.
- Comprehend the applications of Time series analysis.
- Describe Components of Time Series Data
- Difference between Additive model and Multiplicative model

8.2 WHY WE USE TIME SERIES ANALYSIS?

Analysis of time series has emerged as an essential tool for firms seeking to enhance decision-making by using data. Through the analysis of historical trends, companies may gain insights into previous achievements and make informed predictions about future results that are practical and applicable. Time series analysis facilitates the transformation of unprocessed data into valuable

insights that firms may use to enhance their performance and monitor past results.

For instance, merchants may analyze seasonal sales trends in order to adjust their inventory and marketing strategies. Energy firms have the potential to use consumption patterns in order to enhance their production schedule. The applications of AI may also be used for identifying abnormalities, such as a rapid decline in website traffic, which might indicate underlying problems or potential possibilities. Financial institutions use it to promptly react to fluctuations in the stock market. Health care systems need it to promptly evaluate patient risk.

Time series data, as opposed to a collection of statistics, provides a narrative of how business circumstances change and develop over a period of time. This viewpoint enables firms to plan forward, identify problems at an early stage, and take advantage of developing possibilities.

8.3 COMPONENTS OF TIME SERIES DATA

Time series data often consists of distinct components that describe the patterns and dynamics of the data as it evolves over time. Through the examination of these constituents, we may enhance our comprehension of the temporal sequence's behavior and construct more precise models. A time series dataset consists of four primary components:

- Trends
- Seasonality
- Cycles
- Noise

1.Trends refer to longterm patterns or movements in data that show a consistent direction. Trends indicate the overall trajectory of the data, whether it is experiencing growth, decline, or stability over a prolonged duration. Trends represent the extended pattern in the data and may unveil the general progression or regression. For instance, the sales of e-commerce have shown a consistent rising trajectory in the last five years.

2. Seasonality refers to regular patterns that occur within a certain time period, such as daily, weekly, or yearly. Seasonality relates to the consistent and predictable patterns that occur at regular intervals, such as the annual increase in retail activity during the Christmas season. Seasonal components display periodic variations that are consistent in terms of time, direction, and amplitude. As an example, the consumption of power tends to increase significantly during the summer months when individuals activate their air conditioning units.

3.Cycles refer to recurring patterns that are longer than seasonal patterns but shorter than trends. Cycles exemplify oscillations that lack a predetermined duration, such as periods of economic growth and decline. These long-term patterns persist for more than a year and do not exhibit consistent magnitudes or durations. An example of such oscillations is provided by business cycles, which alternate between periods of expansion and downturn.

4. Noise refers to the remaining variability in the data that cannot be accounted for by the other components. Noise refers to the irregular and unexpected variations that remain after accounting for trends, seasonality, and cycles.

Types of Data

Prior to commencing time series analysis, it is often essential to comprehend the nature of the data at hand. The classification may be essentially classified into three unique types: Time Series Data,

Cross-Sectional Data, and Pooled Data. Every kind has distinct characteristics that direct the further examination and modeling.

1. Time series data consists of observations that are gathered at various time periods. Its primary focus is on examining trends, cycles, and other temporal patterns.
2. Cross-sectional data refers to data points that are gathered at a certain instant in time. This is a valuable tool for comprehending the connections or contrasts between distinct items or categories at a particular moment.
3. Pooled data refers to a merged dataset that include both time series and cross-sectional data. This hybrid enhances the dataset, enabling more intricate and extensive studies.

8.4 IMPORTANT TIME SERIES TERMS & CONCEPTS

Time series study is a special field of statistics that examines data points gathered or recorded in a sequential manner over a period of time. It utilizes many approaches and procedures to detect patterns, predict future data points, and make well-informed judgments based on temporal correlations among variables. This analytical methodology utilizes a variety of terminology and ideas that aid in the examination and understanding of data that changes over time.

Interdependence: The correlation between two observations of the same variable at distinct time intervals is essential for comprehending temporal connections.

Stationarity refers to a quality in which the statistical properties, such as the mean and variance, remain constant across time. It is typically a necessary condition for the application of different statistical models.

Differencing is a transformation method used to convert stationary time series data into non-stationary data by subtracting consecutive or delayed values.

Specification: The process of selecting a suitable analytical model for time series analysis may comprise factors such as the curve type or the level of differencing.

Exponential Smoothing is a forecasting technique that use a weighted average of previous observations, giving more importance to more recent data points in order to make short-term forecasts.

Curve fitting refers to the use of mathematical functions to accurately match a collection of data points, often used for data sets that exhibit non-linear connections.

ARIMA (Auto Regressive Integrated Moving Average) is a commonly used statistical model for evaluating and predicting time series data. It incorporates elements such as auto-regression, integration (differencing), and moving average.

8.5 TIME SERIES ANALYSIS TECHNIQUES

Analysis of time series is essential for organizations to forecast future outcomes, evaluate historical achievements, or detect underlying patterns and trends in different measurements. Time series analysis provides significant insights into the fluctuations of stock prices, sales statistics, consumer behavior, and other factors that are dependent on time. By using these methodologies, enterprises may make well-informed choices, streamline processes, and improve long-term goals.

8.6 ADVANTAGES OF TIME SERIES ANALYSIS

Time series analysis provides several advantages to enterprises. The applications of this technology are many and include several areas, such as improving inventory management by accurately predicting sales, strategically planning marketing campaigns by understanding consumer behavior patterns, and devising investment strategies by researching financial markets. Various methodologies serve diverse objectives and provide different levels of detail and precision, underscoring the need for firms to comprehend the approaches that most effectively align with their particular requirements.

Time series analysis has various benefits and may provide significant insights into trends, patterns, and underlying structures present in the data. In this discussion, we will examine many significant benefits of time series analysis:

1. Time series analysis facilitates the identification of trends and patterns present in the data. Through the analysis of consecutive data points, analysts may determine if there is a positive, negative, or consistent trend over a period of time. Identifying these patterns is essential for making well-informed judgments and forecasts.
2. Time series analysis has a key benefit in its capacity to predict future values by using past data. Analysts may use autoregressive integrated moving average (ARIMA) models and exponential smoothing approaches to forecast future trends, enabling firms and governments to engage in proactive planning.
3. Seasonal Analysis: Numerous time series data demonstrate recurring trends that occur at regular intervals, such as significant increases in sales around holidays or variations in

temperature throughout the year. Time series analysis enables the identification and quantification of seasonal impacts, hence facilitating improved planning and allocation of resources.

4. Anomaly Detection: Time series analysis is a powerful method for identifying and flagging unusual or exceptional data points within a dataset. Abrupt increases or decreases in the time series may suggest atypical occurrences or inaccuracies in the data gathering procedure. Detecting such anomalies is crucial for comprehending and resolving inconsistencies in the dataset.
5. Decision Support: Time series analysis is used by businesses and organizations to facilitate decision-making processes. Time series analysis is valuable for making educated and strategic choices in several fields such as finance, inventory management, and resource planning. It involves studying historical patterns and trends.
6. Risk Management: Time series analysis is a powerful tool for evaluating and controlling risks. Organizations may enhance their ability to make educated financial choices and execute risk mitigation methods by anticipating probable future trends and comprehending the volatility in the data.
7. Policy Evaluation: In disciplines such as economics and public policy, time series analysis is used to assess the effects of policies and interventions over a period of time. Through the comparison of data before to and subsequent to the adoption of policies, analysts may evaluate the efficacy of different measures and provide suggestions for enhancements.
8. Optimizing resource allocation involves analyzing the trends in time series data. Efficiently allocating resources to meet future demands is crucial, especially in domains like energy

consumption, where predicting demand patterns plays a significant role.

9. Model Validation: Time series analysis offers a structure for verifying the accuracy of models. Analysts have the ability to assess model predictions against real-world results over a period of time, enhancing and enhancing models for more precise future forecasts.
10. Scientific Research: Time series analysis is used in scientific research to examine phenomena that undergo changes over time, such as climatic data, biological processes, and medical patterns. This enables researchers to get a more profound comprehension of the fundamental dynamics of complex systems.

Time series analysis is a highly adaptable and essential technique that has applications in a wide range of areas. The capacity to reveal patterns, predict forthcoming values, and facilitate decision-making processes makes it a vital element of data analysis in the contemporary day. The benefits of time series analysis, whether in business, finance, or scientific study, enhance decision-making and deepen comprehension of dynamic systems.

8.7 DISADVANTAGES OF TIME SERIES ANALYSIS

Although time series analysis is a useful technique for comprehending and forecasting patterns in sequential data, it does include several drawbacks and constraints. Below are few significant drawbacks associated with time series analysis:

1. The stationarity assumption is a common feature in many time series models, where it is assumed that the statistical characteristics of the time series, such as its mean and

variance, stay constant during the whole time period. In practical use, this assumption may be invalid, resulting in imprecise outcomes.

2. Time series data is susceptible to missing values, outliers, and other data quality problems. Properly managing missing information is essential, as outliers may greatly affect the precision of models.
3. Initial condition sensitivity refers to the tendency of some time series models, particularly those that include iterative processes or recursive forecasting, to be influenced by the starting circumstances. Minor alterations in the initial numbers may result in markedly different predictions.
4. Overfitting: Time series data may display intricate patterns, which increases the possibility of models being excessively tailored to noise present in the data. Overfit models exhibit high performance on historical data but lack the ability to generalize to novel, unknown data.
5. Restricted Forecasting Timeframe: Time series models are often more appropriate for predicting outcomes within a limited time frame, ranging from short to medium term. As the forecasting horizons get longer, the level of uncertainty tends to rise, which may lead to a decrease in the accuracy of forecasts.
6. Seasonality and Non-linearity: Numerous time series approaches presuppose linearity and may encounter difficulties in capturing non-linear patterns or intricate seasonality. Data exhibiting such features may need the use of advanced models.
7. Model selection and complexity: Selecting an appropriate model for a given time series might pose difficulties. It is crucial to carefully evaluate the choice of suitable

parameters and the total complexity of the model, since there is no universally applicable solution.

8. External elements: Time series models sometimes fail to include exogenous elements or events that might have an effect on the data. Traditional time series models may fail to accurately account for economic recessions, policy changes, or unforeseen occurrences.
9. Data Length: Certain time series models need a substantial quantity of past data in order to provide precise forecasts. When there is a scarcity of data, the effectiveness of models may be undermined.
10. Assumption of Independence: Time series models often assume that observations are independent, whereas in actuality, data points may exhibit correlation. Disregarding correlation might result in partial estimations and imprecise forecasts.

Notwithstanding these drawbacks, time series analysis continues to be a potent tool when used correctly, and several strategies have been devised to tackle some of these difficulties. Comprehending the constraints of time series models is crucial for acquiring dependable and significant insights from the data.

8.8 ADDITIVE MODEL:

The additive model in time series analysis is a mathematical structure used to depict the constituent elements that contribute to the overall structure or behavior of a time series dataset. Time series data comprises sequential observations gathered over a period, such as stock prices, temperature measurements, or sales numbers. Gaining insight into and predicting these temporal patterns is essential in several domains, such as finance, economics, and meteorology. The additive model dissects a time series into many

components, with each component reflecting a distinct facet of the data's unpredictability.

An additive model assumes that a time series may be represented as the sum of its distinct components: trend, seasonality, and residual (or error). Now, we will examine each individual element in detail:

1. Trend: The trend component represents the enduring, fundamental pattern in the time series over an extended period. It signifies the overarching trajectory or consistent pattern that endures over a prolonged duration. Trends may have either a linear nature, characterized by a consistent pace of change, or a nonlinear one, displaying more intricate patterns. It is crucial to identify and model the trend in order to comprehend the overall direction of the data.

2. Seasonality is the phenomenon of repeated patterns or variations that adhere to a consistent and predictable schedule. These patterns exhibit periodicity, occurring at regular intervals, such as daily, monthly, or annual cycles. Seasonal impacts may arise due to external variables such as weather conditions, holidays, or cultural events. Integrating seasonality into the model enhances the accuracy of data representation, especially when distinct patterns recur at precise intervals.

3. Residual (Error): The residual component represents the stochastic and unforeseeable variations in the time series that are not accounted for by the trend and seasonality. It quantifies the discrepancy between the actual values and the values anticipated based on the trend and seasonality factors. Examining the residuals is essential for evaluating the precision of the model and detecting any lingering patterns or abnormalities.

The additive model may be mathematically represented as:

$$\mathbf{Y(t)=T(t)+S(t)+R(t)}$$

where:

- Y(t) is the observed value at time t,
- T(t) is the trend component,
- S(t) is the seasonality component, and
- R(t) is the residual (error) component.

The additive model postulates that the impact of the trend, seasonality, and residuals is cumulative, hence facilitating a clear and practical interpretation and implementation. Forecasting future values in an additive model requires estimating the individual components, projecting them into the future, and then adding them together to generate the anticipated values.

The use of additive models offers several benefits in the study of time series:

• Interpretability: The process of breaking down the time series into trend, seasonality, and residual components enhances comprehension of the fundamental patterns that impact the data.

Forecasting involves using the recognized and modeled individual components to produce more precise predictions about future values in the time series.

• Anomaly Detection: By examining the residuals, one might find atypical patterns or outliers that may suggest unforeseen occurrences or data problems.

To summarize, additive models are helpful tools for analyzing time series data and comprehending the fundamental structures that influence its behavior. Through the process of breaking down a time series into its fundamental elements, analysts and researchers may acquire valuable insights, create more accurate forecasts, and deepen their comprehension of the intrinsic dynamics within the data.

Additive models refer to a specific sort of time series forecasting model that breaks down a time series into its distinct components, including trend, seasonality, and residual. Below are the benefits and drawbacks of additive models in the context of time series forecasting:

Benefits:

Explain ability:

dditive models provide a clear and comprehensible depiction of the several elements (trend, seasonality, and residual) inside the time series.

By analyzing each component individually, one may get a clearer comprehension of the underlying patterns present in the data.

Adaptability:

Time series data with diverse patterns and structures may be analyzed using additive models.

They possess the capability to manage both linear and non-linear trends, as well as various forms of seasonality.

User-Friendliness:

Implementing additive models is quite straightforward and requires minimum parameter adjustment.

They are applicable for both immediate and extended prediction assignments.

Resilience:

Additive models have a high degree of resilience against outliers in the data due to their ability to decompose the data and concentrate on the fundamental patterns.

Drawbacks:

Additivity Assumption:

Additive models make the assumption that the different components (trend, seasonality, and residual) are combined by addition. Occasionally, this assumption may be invalid, resulting in imprecise predictions.

Partial acquisition of non-linear patterns:

Additive models may have difficulties in capturing intricate non-linear connections in the data, since they depend on the assumption of linearity in the trend and seasonality components.

Seasonal Patterns Sensitivity:

Additive models may have limited performance when used to time series data that displays severe multiplicative seasonality. For such situations, multiplicative models may be more suitable.

Lack of capacity to capture evolving patterns:

If the time series displays dynamic patterns over time, such as emerging trends or fluctuating seasonality, additive models may struggle to effectively adjust to these changes.

Restricted Predictive Timeframe:

The use of additive models for long-term forecasting may be limited, particularly when the data exhibits substantial changes in underlying patterns.

Reliance on the precision of decomposition:

The precision of the predicting is highly dependent on the precision of the decomposition process. If the decomposition fails to correctly reflect the fundamental constituents, it may undermine the predicting performance.

To summarize, additive models are valuable in many time series forecasting situations, although their effectiveness might be affected by the individual attributes of the data. When selecting a forecasting strategy for a certain time series dataset, it is crucial to take into account the assumptions and limits of additive models.

8.9 MULTIPLICATIVE MODEL

The notion of a multiplicative model is crucial in the study of time series, especially for data that displays patterns and seasonality. This modeling methodology entails breaking down a time series into three primary constituents: trend, seasonality, and randomness or error. Analysts may get significant insights into the underlying patterns and enhance the accuracy of their forecasts by comprehending and modeling these components individually.

1. The trend component of a time series denotes the extended-term progression or orientation of the data. It captures the long-term persistent trend. In a multiplicative model, the trend is seen as a proportional augmentation or reduction over time. Consequently, the trend is not a constant value, but rather varies proportionally with the size of the time series. For instance, if the underlying trend is increasing, then succeeding period's value would be a multiple of the prior one.

2. The seasonality component of a time series refers to the recurring patterns or variations that happen at regular intervals, often linked to certain seasons, months, or days of the week. Seasonality in a multiplicative model is represented as a ratio compared to the trend. This suggests that the impact of the seasonal effect becomes more noticeable as the trend grows stronger. For example, when the trend increases by 10%, the seasonal component will likewise rise by 10%.

3. The residual portion of the time series, which cannot be accounted for by the trend and seasonality, is ascribed to randomness or error. This component captures the non-systematic and unpredictable fluctuations or noise in the data. In a multiplicative approach, randomness is quantified as a fraction of the total magnitude of the time series. As the trend and seasonality rise, the magnitude of the random component likewise grows.

The mathematical FA multiplicative model is a key idea in time series analysis, especially for data that shows patterns and seasonality. This modeling methodology entails breaking down a time series into three primary components: trend, seasonality, and randomness or error. Analysts may get significant insights into the underlying patterns and enhance the accuracy of their forecasts by comprehending and modeling these components individually.

1. The trend component of a time series denotes the extended-term progression or orientation of the data. It captures the long-term, persistent general trend. In a multiplicative model, the trend is seen as a proportional augmentation or diminution over time. Consequently, the trend is not a constant value, but rather varies proportionally with the size of the time series. For instance, if there is a consistent upward trend, each succeeding period's value will be a multiple of the prior one.

2. The seasonality component of a time series refers to the recurring patterns or variations that happen at regular intervals, often linked to certain seasons, months, or days of the week. Seasonality in a multiplicative model is represented as a ratio compared to the trend. This suggests that the impact of the seasonal effect becomes more noticeable as the trend grows stronger. For example, if there is a 10% rise in the trend, the seasonal component will likewise see a corresponding 10% increase.

3. The residual portion of the time series that cannot be accounted for by the trend and seasonality is ascribed to randomness or error. This component captures the non-systematic and unpredictable fluctuations or noise in the data. In a multiplicative approach, randomness is expressed as a fraction of the total magnitude of the time series. As the trend and seasonality rise, the magnitude of the random component likewise grows.

Mathematical Expression: A time series that exhibits multiplicative behavior.

Yt can be represented as the product of its trend Tt, seasonality St, and randomness Et:

$Yt=Tt\times St\times Et$

where:

- Yt is the observed value at time t,
- Tt is the trend component at time t,
- St is the seasonality component at time t,
- Et is the randomness or error component at time t.

To summarize, a multiplicative model in time series analysis is an effective tool for comprehending and forecasting intricate patterns in data. Through the process of decomposing the time series into its constituent parts of trend, seasonality, and randomness, analysts are able to construct more precise models and make well-informed judgments by gaining a comprehensive grasp of the fundamental dynamics at play.

The multiplicative model is often used in time series analysis, especially for data that has a fluctuating trend or seasonality. Similar to other modeling approaches, the multiplicative model has its own set of benefits and drawbacks.

Benefits:

The multiplicative model is very adaptable and suitable for time series data that exhibits both trend and seasonality. It enables the variable depiction of both components, making it suitable for a diverse variety of real-world situations.

Reflecting changing patterns: Multiplicative models are useful for analyzing data that exhibits varying amplitudes of seasonal fluctuations or changing strengths of trends over time. Multiplicative models, unlike additive models, are capable of capturing dynamic patterns more effectively by not assuming constant seasonal amplitudes.

Optimal Choice for Exponential Growth: When the underlying data demonstrates exponential growth or decay, the multiplicative model is the most suitable option. It is applicable in cases when the percentage change in the series is directly proportionate to its present level, making it useful for economic, demographic, or technological adoption statistics.

Grips Non-Constant Variability: In some time series, the data's variability may exhibit fluctuations in its magnitude, either increasing or decreasing with time. Multiplicative models may accommodate the non-constant variability, so offering a more precise depiction of the underlying patterns.

Statistical Interpretability: The multiplicative decomposition facilitates a lucid explanation of the distinct components, including trend, seasonality, and residuals. The ability to comprehend the time series behavior is very helpful in understanding its underlying factors.

Drawbacks:

Sensitivity to Outliers: Multiplicative models are highly responsive to outliers, which are extreme values that differ greatly from the main trend. Outliers may have a disproportionate influence on the accuracy of multiplicative models, perhaps leading to the misunderstanding of trends and seasonality.

The multiplicative model's high flexibility may result in over fitting, particularly when used to datasets containing noise or irregular patterns. Over fitting arises when the model incorporates the irrelevant details included in the data, hence reducing its ability to accurately predict future observations.

Stationarity Assumption: Multiplicative models assume stationarity, which implies that the statistical characteristics of the time series remain constant over its duration. Practically, several time series in the real world have non-stationary characteristics, and using a multiplicative model without accounting for this might lead to distorted estimations.

Interpretation Challenge: Although the multiplicative model offers distinct components for interpretation, non-experts may find it difficult to comprehend and interpret the multiplicative decomposition. The intricacy of the situation might impede the effectiveness of communication and decision-making processes.

Data Transformation Requirement: In some instances, it may be imperative to convert the data in order to fulfill the assumptions of the multiplicative model. For instance, it may be necessary to apply logarithmic transformation to the data in order to achieve stability, which might introduce complexity in the interpretation of the findings.

Ultimately, the selection between multiplicative and additive models relies on the specific attributes of the time series data and

the goals of the research. Although the multiplicative model provides flexibility in capturing dynamic patterns, it is important to acknowledge its limits and possible difficulties, particularly when handling outliers and non-stationary data.

8.10 DIFFERENCE BETWEEN ADDITIVE MODEL AND MULTIPLICATIVE MODEL

1. Combination method:

The components are combined through addition in the Additive Model.

The multiplicative paradigm entails the multiplication of components.

2. Seasonal fluctuations:

Additive Model: The seasonal fluctuation remains consistent over time.

The multiplicative model posits that the seasonal variation of a series is directly proportional to the series' level.

3. Influence of the Trend:

The additive model illustrates that the series is linearly affected by the trend.

The multiplicative model posits that the series is proportionally affected by the trend.

4. Possible scenarios:

Additive Model: Suitable for time series data that demonstrate consistent seasonal fluctuations.

The multiplicative model is suitable for time series data that demonstrate proportionate seasonal fluctuations.

It is imperative to comprehend the characteristics of time series data in order to select the appropriate model, as this decision affects the interpretation and prediction of the components.

8.11 LET US SUM UP :

Time series analysis is the examination of data points that are gathered or recorded at regular time periods. The main goal is to comprehend the fundamental structure and mechanism that generated the observations and use this comprehension to predict future values. Below is a summary of fundamental principles and techniques in time series analysis:

Core principles

Trend refers to the overall pattern or direction that may be seen in the data over a significant period of time.

Seasonality refers to the occurrence of regular and recurring patterns or cycles in data that are associated with certain time periods, such as monthly or annual intervals.

Noise refers to the presence of random fluctuations or variations in the data.

Stationarity refers to the property of a time series where its statistical characteristics, such as the mean and variance, remain constant and do not change over time.

Approaches and Strategies

Decomposition refers to the process of dividing a time series into its constituent parts, namely the trend, seasonal, and residual components.

Additive decomposition is a method that assumes the components

of a time series add up. This approach is particularly beneficial when the variances around the trend do not change based on the level of the time series.

The multiplicative decomposition assumes that the components of the time series multiply together. This is particularly beneficial when the fluctuations around the trend rise as the level of the time series increases.

Smoothing: Methods used to eliminate unwanted disturbances and emphasize patterns.

Moving Average: Calculates the mean of data points within a defined frame in order to reduce fluctuations in the series. The additive model in time series analysis is a mathematical framework used to represent the individual components that contribute to the overall structure or behavior of a time series dataset. Time series data consists of consecutive observations collected over a duration, such as stock prices, temperature readings, or sales figures. Acquiring understanding and forecasting these time-based patterns is crucial in several fields, including finance, economics, and meteorology. The additive model decomposes a time series into many components, where each component represents a unique aspect of the data's unpredictability.

The concept of a multiplicative model is essential in the analysis of time series, particularly for data that exhibits patterns and seasonality. This modeling approach involves decomposing a time series into three main components: trend, seasonality, and random variation or error. By independently analyzing and modeling these components, analysts may get valuable insights into the underlying patterns and improve the accuracy of their projections.

8.12 KEY WORDS :

Trend: The extended period of time during which the data consistently moves or progresses in a certain direction.

Seasonality: refers to the occurrence of regular and repetitive patterns or cycles in data that are associated with certain time periods, such as monthly or annual intervals.

Noise: refers to the presence of random fluctuations or variations in the data.

Stationarity: refers to a time series that maintains consistent statistical features, such as a steady mean and variance, throughout its duration.

8.13 ANSWERS TO CHECK YOUR PROGRESS

1. A …… is a succession of data points that are usually measured at consecutive time intervals.

2. The primary constituents of a time series consist of ……components.

3. The ……component signifies the extended-term evolution of the series.

4. In an ……..model, the value of a time series at any given moment is represented as the total of its individual components.

5. …….refers to the process of recognizing and eliminating the influence of seasonal patterns from a time series.

6. In a ……..model, the value of a time series at any given moment is calculated by multiplying its individual components together.

7. The use of a multiplicative model is common when the …….is directly proportional to the overall level of the series.

8. In a multiplicative model, the rise in the trend component leads to a proportional ……in the seasonal variations.

Answer:

1. time series
2. trend, seasonal, cyclical, and irregular
3. trend
4. additive
5. Deseasonalizing
6. Multiplicative
7. seasonal fluctuation
8. increase

8.14 TERMINAL QUESTIONS

Q1. Explain concept of Time series analysis.
Q2. What are the components of Time series?
Q3. Distinguish between Additive model and Multiplicative model.
Q4. Discuss causes of variations in the time series data.
Q5.Discuss role of time series in business.

UNIT 9: SEASONAL VARIATION, CYCLICAL VARIATION

Structure

9.0 Objectives

9.1 Introduction

9.2 Seasonal Variation Characteristics

9.3 Determining Seasonal Fluctuation

9.4 Quantifying Seasonal Fluctuation

9.5 Cyclical variation

9.6 Attributes of Cyclical Variation

9.7 Factors contributing to cyclical variation

9.8 Techniques for Detecting Cyclical Fluctuations

9.9 Consequences of Cyclical Variation

9.10 Difficulties in analyzing Cyclical Variation

9.11 Difference between Seasonal variation and cyclical variation

9.12 Let Us Sum Up

9.13 Key Words

9.14 Answers to Check Your Progress

9.15 Terminal Questions

9.0 OBJECTIVES

After studying this unit, you should be able to:

- Define seasonal variation and its importance in time series analysis.
- Determine the factors that contribute to seasonal fluctuations in various scenarios.
- Acquire techniques for recognising and quantifying seasonal fluctuations.

- Define cyclical variation and distinguish it from seasonal variation

9.1 INTRODUCTION

Seasonal variation in time series analysis is an essential notion for comprehending and forecasting trends in data that display consistent swings across certain time periods. It is crucial, especially in disciplines like economics, finance, meteorology, and several others, where data points are gathered chronologically. This article explores the characteristics of seasonal variation, approaches for recognizing and quantifying it, and strategies for integrating it into time series models.

Seasonal variation pertains to the cyclic variations that often occur in time series data as a result of seasonal variables. These fluctuations might be seen in daily, weekly, monthly, or annual patterns. For instance, retail sales often see a significant increase during the holiday season, power consumption rises throughout the summer and winter months as a result of heating and cooling requirements, and agricultural outputs fluctuate in accordance with planting and harvest seasons.

The primary attribute of seasonal variation is its inherent predictability; the pattern recurs consistently throughout a predetermined time. Seasonal variation may be distinguished from other forms of fluctuations in time series data, such as trends or random noise, by their predictability.

9.2 SEASONAL VARIATION CHARACTERISTICS

Regular Intervals: Seasonal fluctuations occur annually at the same time. For instance, the holiday season may result in an increase in retail sales each December.

Predictability: These variations are predictable in accordance with historical data. Businesses are capable of anticipating these modifications due to their frequency.

Recurrence: The seasonal variation pattern is repeated annually. For instance, the sales of air conditioners experience an increase during the summer and a decrease during the winter.

Cause: They are frequently precipitated by cultural events, holidays, and fluctuations in the weather. For example, agricultural yields may fluctuate in accordance with the growing season.

9.3 DETERMINING SEASONAL FLUCTUATION

Detecting seasonal variation entails examining time series data to identify recurring trends. Typical techniques used for this identification comprise:

Visual inspection is the act of representing the data on a graph, which may often expose discernible patterns. Seasonal impacts manifest as recurring peaks and valleys in the data at reliable periods.

The Autocorrelation Function (ACF) quantifies the correlation between a time series and a delayed version of itself over different

time periods. Noticeable increases at certain time intervals suggest the existence of recurring patterns. For example, in monthly data, a sudden increase at a lag of 12 indicates a yearly seasonal trend. The Seasonal Decomposition of Time Series (STL) method breaks down the series into three distinct components: trend, seasonality, and residuals. STL facilitates the isolation and comprehension of the seasonal component by decomposing the series into its additive or multiplicative constituents.

9.4 QUANTIFYING SEASONAL FLUCTUATION

Accurate modeling and forecasting need the use of several methodologies to quantify the seasonal component. Several of these techniques comprise:

The Classical Decomposition approach entails the process of dissecting the series into its distinct components, namely the trend, seasonal, and irregular components. The seasonal component is determined by calculating the mean of the data points that correspond to each specific time, such as each month for monthly data.

Centered moving averages are used to reduce short-term swings and highlight the seasonal trend. For example, using a 12-month centered moving average on monthly data might assist in detecting yearly patterns.

The Fourier Transform is a technique that translates time series data into the frequency domain. The peaks seen in the frequency spectrum are indicative of the prevailing seasonal patterns present in the data.

Integrating Seasonal Fluctuations into Time Series Models After being found and quantified, the seasonal component has to be included into time series models in order to enhance the accuracy of forecasting. Multiple methodologies may be employed:

Seasonal dummy variables are used in regression models to indicate seasonal impacts. For instance, while analyzing monthly data, it is possible to use eleven dummy variables to represent the impact of each month, while designating one month as the baseline. Seasonal ARIMA (SARIMA) models: The ARIMA model, short for AutoRegressive Integrated Moving Average, is a very popular method used in the analysis of time series data. The SARIMA model has seasonal components in addition to non-seasonal components, allowing it to capture both seasonal and non-seasonal fluctuations. The model is represented by the notation ARIMA(p,d,q)(P,D,Q)s, where (P,D,Q)s indicate the seasonal component, with s being the length of the seasonal period.

Exponential Smoothing State Space Models (ETS) are an extension of exponential smoothing approaches that explicitly consider seasonality. The ETS framework has error, trend, and seasonality components, enabling versatile modeling of different time series patterns.

The process of seasonal decomposition and forecasting entails breaking down the series into its trend, seasonal, and residual components. Each component is then forecasted individually, and then combined to get the ultimate prediction. Decomposition is often performed using methods such as STL and X-13ARIMA-SEATS. Recent developments in machine learning have brought to the introduction of models like Seasonal Neural Networks and Seasonal-LSTM networks. These models have the ability to

automatically identify and understand seasonal patterns in data, without the need for human dissection.

Pragmatic Factors

Although the inclusion of seasonal fluctuation enhances the accuracy of the model, it also increases a level of complexity. Several pragmatic factors to consider are:

Over fitting occurs when a model has an excessive number of seasonal components or too complicated models, resulting in good performance on previous data but poor performance on fresh data. Regularization methods and cross-validation may assist in reducing this potential danger.

Seasonality may undergo changes over time as a result of external causes such as economic fluctuations, climatic change, or alterations in consumer behavior. It is necessary to regularly update models in order to correctly represent these changes.

several Seasonalities: Occasionally, data may display several seasonal trends. For instance, the daily sales data may exhibit both weekly and annual seasonal patterns. Advanced methodologies such as TBATS, which incorporates Trigonometric, Box-Cox transformation, ARMA errors, Trend, and Seasonal components, are very successful in managing numerous seasonal patterns. Time series data often includes missing values or anomalies. Preprocessing techniques, such as interpolation to handle missing data and robust approaches to identify outliers, are crucial for achieving accurate seasonal modeling.

In conclusion

Seasonal variation is a basic component of time series analysis, which represents the regular and predictable oscillations that occur within specified time periods. It is essential to identify, measure, and include seasonal influences in time series models in order to achieve reliable forecasting. Analysts have a variety of strategies at their disposal, ranging from classical decomposition to recent machine learning approaches, to effectively address seasonality.

Nevertheless, it is crucial to thoroughly analyze model complexity, evolving patterns, various seasonalities, and data quality to guarantee resilient and dependable forecasting. Through the acquisition of these skills, analysts may get a more profound understanding of their data and thus make more well-informed judgments by relying on precise forecasts.

9.5 CYCLICAL VARIATION

Cyclical variation in time series refers to the extended variations or oscillations that occur around the trend line of a time series. Contrary to seasonal variation, which happens at consistent and predictable intervals throughout a year, cyclical variation does not adhere to a certain timeframe. However, it is impacted by wider economic, social, or ecological factors that result in cycles of different durations. Comprehending cyclical variance is essential for precise prediction, economic strategizing, and strategic decision-making.

9.6 ATTRIBUTES OF CYCLICAL VARIATION

Temporal and spatial extent: Cyclical fluctuations occur across extended periods of time and are not limited to a particular timeframe. The duration of these cycles may exhibit significant variation, often extending over many years or even decades. Amplitude refers to the amplitude of cyclical variations, which may vary depending on the intensity and length of the underlying variables that cause the cycle. For example, economic cycles may see significant declines during recessions and substantial increases during booms.

Direction: Cyclical fluctuations demonstrate alternating periods of expansion (moving upward) and contraction (moving below). The phases may exhibit either symmetry or asymmetry, contingent upon the underlying reasons.

9.7 FACTORS CONTRIBUTING TO CYCLICAL VARIATION

Cyclical changes emerge from a multitude of variables, encompassing:

Economic factors, such as business cycles, are the main contributors to cyclical fluctuations in the economy. These include periods of economic growth, climax, economic decline, and subsequent improvement. Economic metrics such as GDP, unemployment rates, and industrial output often display cyclical patterns.

technical Changes: Innovations and technical progress may result in alternating periods of expansion and contraction. For example, the

use of novel technology may enhance productivity and stimulate economic expansion, but the obsolescence of outdated technologies might result in economic downturns.

Political and social factors such as changes in government policy, political stability, and societal trends may have an impact on cyclical fluctuations. Political events such as elections, wars, and major policy changes may lead to substantial cyclical oscillations. Natural Phenomena: Environmental and climatic fluctuations may induce periodic variations in agriculture productivity and other businesses reliant on natural resources. For instance, long-term weather patterns such as El Niño and La Niña may have a substantial influence on agricultural production and economic outcomes.

9.8 TECHNIQUES FOR DETECTING CYCLICAL FLUCTUATIONS

Visual examination: Analyzing the time series data by plotting it and visually examining the graph might facilitate the detection of cyclical patterns. Peaks and troughs may be studied to ascertain the existence and regularity of cycles.

Moving averages, such as smoothing methods, may effectively emphasize cyclical elements by minimizing short-term variations. By calculating the mean of data points within a certain time frame, the underlying patterns or cycles become more evident.

Decomposition Methods: Time series decomposition involves the separation of a series into its constituent components, namely trend, seasonal, and residual. The use of additive and multiplicative decomposition techniques may effectively separate the cyclical component from the whole series.

Spectral analysis is a method that includes converting time series data into the frequency domain using Fourier transform. Spectral analysis is a method used to identify the most prominent cycles and their corresponding frequencies in a series. This analysis helps to get a better understanding of the cyclical patterns shown by the series. Econometric models, such as the Hodrick-Prescott (HP) filter and Baxter-King filter, are used to isolate the cyclical component of a time series. These models aid in the process of reducing the variability in a series and separating the cyclical patterns from the long-term trends and random fluctuations.

9.9 CONSEQUENCES OF CYCLICAL VARIATION

Comprehending cyclical variation has several significant consequences:

Economic forecasting relies on the identification and analysis of cyclical patterns, which are crucial for achieving accurate predictions. Policymakers and companies use this data to forecast economic contractions and expansions, enabling them to make well-informed choices.

Investment Strategies: Investors use cyclical analysis to strategically manage their investments, purchasing assets during periods of economic decline and divesting them during periods of high performance. Gaining insight into market cycles is crucial for formulating efficient investing strategies and mitigating risks. Business Planning: Enterprises may use their understanding of cyclical fluctuations to improve their operations, effectively manage inventories, and strategically plan production schedules. This facilitates the synchronization of corporate operations with anticipated economic circumstances.

Policy Formulation: Governments and central banks use cyclical analysis to create and execute economic strategies. Comprehending the different stages of economic cycles facilitates the formulation of counter-cyclical strategies to maintain stability in the economy.

9.10 DIFFICULTIES IN ANALYZING CYCLICAL VARIATION

Non-stationarity refers to the common occurrence of time series data displaying patterns that change over time, which poses difficulties in identifying and analyzing cyclical patterns. Non-stationary data exhibit fluctuations in both their average and variability across time, making it more challenging to identify cycles.

Constraints in data: Insufficient availability and subpar quality of data might impede the precise detection of cyclical fluctuations. Comprehensive data over a long period of time is sometimes necessary to accurately capture whole cycles, and the absence or presence of inaccurate data might affect the study.

Complex Interactions: Cyclical changes are impacted by a plethora of elements, making it challenging to separate the effect of particular variables. The study is made more difficult by the interactions between economic, social, and environmental elements. The type and duration of cycles might undergo alterations throughout time as a result of structural changes in the economy or other fundamental variables. Using past data to correctly anticipate future cycles becomes difficult due to this.

In conclusion, cyclical variation in time series is an essential aspect of analyzing long-term data. It represents the larger economic, social, and ecological cycles that impact the series. It is crucial to identify and comprehend these changes in order to make precise

forecasts, develop strategic plans, and create efficient policies. Although there are difficulties in studying cyclical fluctuations, progress in statistical and econometric techniques offers significant instruments for discovering and comprehending these patterns. Therefore, cyclical analysis continues to be a crucial component of time series analysis and economic study.

9.11 DIFFERENCE BETWEEN SEASONAL VARIATION AND CYCLICAL VARIATION

The characteristics and causes of seasonal and cyclical variations are distinct, despite the fact that they are both categories of fluctuations observed in time series data:

Definition: Seasonal variation is the term used to describe periodic fluctuations that occur at regular intervals as a result of seasonal factors.in the event that Cyclical variation is a term that denotes fluctuations that persist for extended periods, typically spanning several years, and are associated with broader economic or business cycles.

Frequency: Seasonal fluctuations typically occur within a single year and recur annually. Whereas The completion of a cycle can require several years, as cyclical variations are not as consistent as seasonal variations.

Causes: Factors such as weather, holidays, and school schedules frequently contribute to seasonal fluctuations. For instance, during the holiday season in December, retail sales may reach their highest point, or during the summer months, ice cream sales may increase. Economic factors, such as business cycles, which encompass periods of expansion and contraction in the economy, frequently

influence cyclical variations. For instance, housing markets or stock prices may undergo cycles of peaks and crashes.

Pattern: The seasonal variation is consistent and predictable over time. The pattern of cyclical variation is less predictable and can be influenced by a variety of external economic factors.

Pattern:

Time Frame: Seasonal variations are found within a single year and occur annually, while cyclical variations are found over multiple years.

Predictability: Seasonal fluctuations are more regular and predictable, while cyclical fluctuations are less predictable and can be influenced by a variety of factors.

Cause: Cyclical variations are influenced by broader economic or business cycles, while seasonal variations are frequently the result of natural or recurring events.

9.12 LET US SUM UP:

Fluctuation in different seasons

Seasonal variation refers to the anticipated and consistent changes in a variable that occur cyclically during the course of a year.

Seasonal fluctuation is mostly influenced by environmental variables such as temperature, daylight duration, and weather patterns.

Industries, such as retail, exhibit seasonal fluctuations, with sales often rising during the Christmas season and declining thereafter.

Seasonal indices are used to quantify and compensate for the influence of seasonal fluctuations in data analysis.

Notable instances of seasonal change include a rise in ice cream purchases during the summer and an increase in heating expenses during the winter.

Cyclical variation refers to the repetitive pattern or fluctuation of a system or phenomenon across time.

Cyclical variation refers to the oscillations in data that transpire at periodic periods, often as a consequence of economic circumstances or business cycles.

Macro-economic metrics such as GDP, inflation rates, or unemployment rates may be used to identify cyclical fluctuations.

Cyclical variation, in contrast to seasonal variation, lacks a predetermined calendar rhythm and may span from a few years to many decades in duration.

Companies often conduct analyses of economic cycles in order to forecast and prepare for probable fluctuations in demand and supply, taking into account projected economic circumstances.

Cyclical variation is a component that is examined in time series analysis, together with trend and seasonal variations, to provide insight into the overall pattern of the data.

9.13 KEY WORDS:

Seasonal variation: is mostly caused by environmental factors, such as fluctuations in temperature, daylight duration, and weather patterns.

Cyclical variation: is often overlaid on the trend and seasonal components in order to assess its influence.

9.14 ANSWERS TO CHECK YOUR PROGRESS:

1. The ……..component represents the predictable fluctuations that occur within a certain time period, such as a year or a quarter.

2. The …….component refers to the stochastic fluctuation that persists even after considering the trend, seasonality, and cyclical components.

3. Seasonal variation refers to the anticipated and consistent oscillations in a factor that happen cyclically…….

4. A ……refers to a recurring pattern of fluctuations in data that takes place at consistent periods beyond one year, often influenced by economic circumstances.

5. ……..variation is often characterized by a lower level of predictability compared to seasonal variation. It may manifest throughout a range of time periods, spanning from a few years to many decades.

Answer:

1. Seasonal
2. Irregular
3. over a year
4. cycle
5. Cyclical

9.15 TERMINAL QUESTIONS:

Q1. Explain seasonal variations and cyclical fluctuation.

Q2. Distinguish between seasonal variations and cyclical fluctuation.

Q3. Explain difficulties faced in analysing Cyclical Variation.

Q4. How determining Seasonal Fluctuations?

Refe

UNIT 10: VARIOUS METHODS OF TIME SERIES ANALYSIS AND THEIR APPLICATIONS IN BUSINESS.

Structure

10.0 Objectives

10.1 Introduction

10.2 Measurement of Trend

10.3 Free-hand Curve Method

10.4 Semi-average Method

10.5 Moving-average Method

10.6 Method of Least Squares

10.7 Short-Term Fluctuation Measurement

10.8 Let Us Sum Up

10.9 Key Words

10.10 Answers to Check Your Progress

10.11 Terminal Questions

10.0 OBJECTIVES

After studying this unit, you should be able to

- Acquire a comprehensive understanding of the basic principles behind time series data.
- Acquire methods for identifying and analyzing trends.
- Investigate smoothing techniques such as moving averages and least squares.
- Comprehend the significance and practical uses of time series analysis in the business context.

10.1 INTRODUCTION

The most often used techniques in time series analysis. 1. Analysis that provides detailed and accurate descriptions. Descriptive analysis offers a foundational comprehension of the features of the data. This approach encompasses:

Trend Analysis: Determines the sustained upward or downward movement in the data over an extended period of time. Seasonal Analysis: Identifies recurring trends within a predetermined time frame, for as on a daily, monthly, or annual basis.

Cyclic Patterns: Identifies variations that do not follow a consistent time frame but instead happen at unpredictable intervals. Irregular/Random Components: Identifies stochastic disturbances or abnormalities that deviate from a predicted pattern.

2. Techniques for achieving smoothness

Smoothing techniques are used to eliminate noise and emphasize trends and patterns.

Typical methods include:

Moving Averages: Calculates the mean of the data for a certain number of time periods in order to reduce the impact of temporary changes.

Simple Moving Average (SMA) is a statistical method that calculates the average of a dataset over a predetermined number of time periods.

The Weighted Moving Average (WMA) is a method that assigns varying weights to distinct data points, with a greater emphasis on more recent observations.

The Exponential Moving Average (EMA) is a mathematical technique that assigns more importance to current data while still taking into account older data by applying exponential decay to the weights.

Exponential Smoothing is a sophisticated approach that involves the exponential lowering of weights. Some of the methods include: Single Exponential Smoothing is appropriate for data that does not have a trend or seasonality.

Double Exponential Smoothing, also known as Holt's Linear Trend Model, is a technique that enhances single smoothing by including trend data.

Triple Exponential Smoothing, often known as the Holt-Winters Method, is a technique that may be used to analyze data that exhibits both trend and seasonality.

3. Methods of breaking down or analyzing anything into its constituent parts Decomposition methods analyze time series data by separating it into many components:

Additive Decomposition: This method assumes that the trend, seasonality, and residuals combine together to create the time series. Multiplicative Decomposition: This method assumes that the different components of the time series interact with each other by multiplying together.

Decomposition facilitates the comprehension and modeling of individual components in isolation.

4. The Autoregressive Integrated Moving Average (ARIMA)

model.
ARIMA models are extensively used for time series forecasting. ARIMA integrates:

The Autoregressive (AR) Model utilizes the relationship between one observation and many previous observations (p) that occurred at different time intervals.

The Integrated (I) Model involves applying differencing to the observations in order to achieve stationarity in the time series (d). The Moving Average (MA) Model utilizes the relationship between an observation and a residual error derived from a moving average model that is applied to previous observations (q).

The ARIMA model is often denoted as ARIMA(p, d, q), with p representing the autoregressive order, d representing the differencing order, and q representing the moving average order. p represents the quantity of lag data that are included into the model. d is the frequency at which the original observations are subtracted from each other.

The variable q represents the magnitude of the moving average window.

5. Seasonal Autoregressive Integrated Moving Average (SARIMA)
SARIMA is an extension of ARIMA that takes into consideration the seasonal patterns in the data. The model includes seasonal features in the autoregressive and moving average components, by including variables for seasonal autoregression, seasonal differencing, and seasonal moving average. The model is often denoted as SARIMA(p, d, q)(P, D, Q, s), where:

The seasonal components are represented by P, D, and Q. s represents the duration of the seasonal cycle.

6. Autoregressive Conditional Heteroskedasticity (ARCH) and Generalized ARCH (GARCH) models.

ARCH and GARCH models are used for analyzing time series data that exhibit periods of high volatility. These models are very valuable in financial time series analysis, since they can effectively capture the phenomenon of volatility clustering that often occurs. An ARCH (Autoregressive Conditional Heteroscedasticity) model is a statistical model that characterizes the variability of the present error term based on the magnitudes of the error terms from prior time periods.

The GARCH model is an extension of the ARCH model that incorporates the prior variances and mistakes in order to describe the variance.

7. VAR (Vector Autoregression)

VAR models are used to analyze the mutual effect of several time series. This technique captures the linear relationships between several time series. Every variable in the system is a linear function of its own past values and the past values of all other variables in the system.

8. Models of State Space

State space models provide a versatile framework for modeling time series data. They are composed of:

The observation equation establishes a relationship between the observed data and the state variables.

The state equation is a mathematical representation that describes the temporal evolution of the state variables.
Kalman filters are often used to estimate the latent state variables in these models. State space models are very flexible and have the ability to include ARIMA models as well as more intricate structures.

9. Techniques for Machine Learning
Time series analysis is seeing a growing use of machine learning methods. Several noteworthy techniques include:
Regression trees and random forests are versatile algorithms that may be used to both regression and classification challenges. Support Vector Machines (SVM) are a kind of machine learning algorithm that are used for classification and regression tasks. SVMs are based on the concept of finding the optimal hyperplane that separates different classes of data points in a high-dimensional space. They are particularly effective in dealing Applicable for both categorization and numerical prediction tasks.

Artificial neural networks: Specifically, Recurrent Neural Networks (RNNs) and extended Short-Term Memory (LSTM) networks are specifically intended to process sequential input and record relationships that span over extended periods of time.

10. Analysis using Fourier transforms
Fourier analysis breaks down a time series into sinusoidal components. It is very beneficial for detecting recurring patterns and comprehending the frequency distribution of the data.

Time series analysis is a multifaceted and intricate discipline that provides a broad range of techniques for comprehending, modeling, and predicting data that varies over time. From fundamental

descriptive approaches to advanced machine learning models, each method has unique characteristics and is suitable for certain applications. The selection of the methodology is contingent upon the characteristics of the data, the inherent patterns, and the particular goals of the research.

10.2 MEASUREMENT OF TREND

The following are the four important methods which are used in estimating the trend:-

(I) Free-hand Curve Method

(III) Moving-average Method

(II) Semi-average Method

(IV) Method of Least Squares

10.3 FREE-HAND CURVE METHOD

'Graphic Method' or 'Curve fitting by inspection' are additional terms that are attached to this approach. The following is the procedure for determining the trend using this method:- Initially, the original values of a time series are plotted on a graph paper, and a histogram is generated by connecting these points.

(2) Subsequently, a uniform curve is plotted through these points, taking into account the direction of fluctuations, in order to ensure that the curve accurately represents the overall trend of the data. Advantages: (1) Simple approach: It is the most straightforward method of trend analysis. There is no requirement for any mathematical calculations, which results in a reduction of time and labor.

(2) Flexible method- It is flexible in the sense that it can be employed to represent both linear and non-linear trends. Advantages: (1) Subjective method: This method is highly subjective due to the potential for different curves to be drawn by different individuals for the same set of data. In other words, the curve may be influenced by the personal bias and judgment of the curve drawer.

(2) Inaccuracy—This approach is not founded on mathematical calculations. Therefore, it is inaccurate.

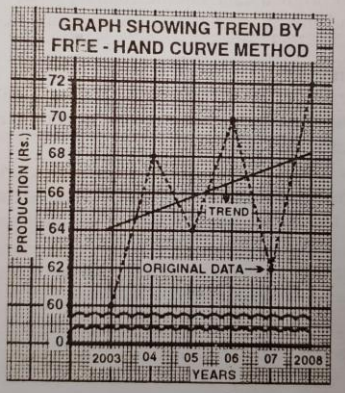
(3) Forecasting Risk-The curve's subjective nature renders it hazardous for forecasting or prediction purposes. The conclusion is that this method is not substantially reliable and scientific, despite its simplicity and adaptability. Therefore, its practical application is exceedingly restricted.

Example :

Show the secular trend from the following data by free-hand curve method:—

Year	:	2003	2004	2005	2006	2007	2008
Production (Rs.)	:	60	68	64	70	62	72

Solution:



10.4 SEMI-AVERAGE METHOD

The process of identifying a trend using the semi-average method is as follows:-

(1) Dividing a time series into two equal portions Initially, the values of time series are divided into two equal portions. For instance, if six years of values are provided, the first three years will be retained in the first section, while the remaining three years will be included in the second section. If the number of values is even, they will be divided precisely into two equal parts. However, if the number of values is odd, the median value (value of the middle year) is omitted, and the remaining values are divided into two equal parts. For instance, if the values of 11 years are provided, the (11+1) /2 = 6 year will be omitted, and the series will be divided into two equal parts based on the values from the first year to the fifth year and the seventh year to the eleventh year.

(2) Calculation of two averages After dividing the given series into two equal portions, we subsequently calculate the arithmetic mean of time-series values for each half separately. Semi-averages are the term used to describe these values. Conversely, the median may be computed.

(3) The original data is plotted on graph paper, and a curve is drawn based on the plot. This is done after the mathematical calculations have been completed.

(4) The semi-averages are depicted as points against the midpoint of the respective time periods spanned by each part. For instance, the first point will be plotted against the median point of the first five years, which is: (5+1)/ 2 = 3^{rd} year and the second point in relation

to the last five years, which is the eighth year. The third

(5) Trend line: The trend of the data is represented by a straight line that is drawn by connecting the two nodes of semi-averages.

Benefits of the Semi-average Method:

(1) Simplicity: This method is straightforward to comprehend and implement, as it involves no mathematical calculations beyond determining the arithmetic mean of two sections of the series.

(2) Two semi-averages are used to establish the Objectivity-Trend line. Therefore, the trend line is characterized by objectivity and certainty. In other words, any two individuals will derive the same trend line from a collection of figures.

(3) Estimates of the past or future-The trend line can be extended on either side to derive estimates of the past or future.

Limitations of the Semi-average technique:

(1) Linear Trend: This technique is suitable for use only when the plotted elements exhibit a linear or approximately linear relationship.

(2) Affect of extreme values-In the event that there are specific extreme values (extremely low or extremely high), the value of semi-averages will be influenced, and the trend line may not accurately represent the values.

It is evident that the semi-average method is more objective than the free hand curve method; however, it is not as reliable, particularly when there are certain extreme values or a lack of linear relationship.

Example:

Determine trend of the following data by applying the method of semi-average and estimate the value for 2010:—

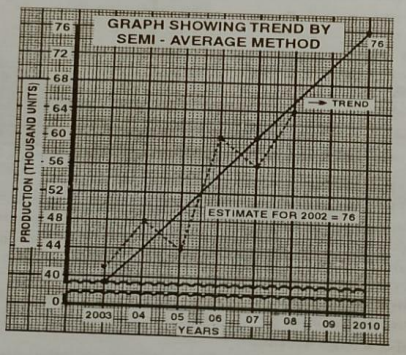
Year	:	2003	2004	2005	2006	2007	2008
Production (thousand units)	:	40	48	44	60	56	64

Solution:

There are six years figures in this question, so they can be divided into two halves of 3 years each 2003 to 2005 and 2006 to 2008. The trend line will be drawn on graph paper on the basis of averages of these two halves.

Year	Production (thousand units)	3 yearly Semi-totals	Semi-averages
2003	40		
2004	48	132	$\frac{132}{3} = 44$
2005	44		
2006	60		
2007	56	180	$\frac{180}{3} = 60$
2008	64		

The semi-average of the first half 44 will be plotted against the mid-point of the first half *i.e.*, the year 2004 and the semi-average of the second half 60 would similarly be plotted against the year 2007.



10.5 MOVING-AVERAGE METHOD

The moving average method is a straightforward and adaptable too l that can be used to reduce fluctuations and acquire trend values w ith a moderate level of precision.

It involves the acquisition of a sequence of moving averages (arith metic means) of successive contiguous groups or sections of the ti me series. For instance, the three-

year moving average is to be calculated for six years, namely a, b, c, d, e, and *f*.

The procedure will be as follows:- (a+b+c)/3, (b+c+d)/3, (c+d+e)/3 , (d+e+f)/3.

The fundamental inquiry that must be resolved in this approach is t he appropriate period for the moving average, such as three, four, o

r five years.

This determination is predicated on the volume of data and its fluct uations.

The topics can be categorized into two groups from the perspective of moving average calculation:--

(1) when the period is odd, and (2) when the period is even.

(1) Moving averages with odd periods-

It refers to the moving averages of odd periods or years, such as 3, 5, 7, 9, 11, and so forth.

The procedure can be elucidated as follows, assuming that three-year moving averages are to be computed:- Initially, three-year moving totals will be obtained.

The sum of the first three years will be compared to the middle of t he three years, which is the second year.

(ii) Subsequently, the sum of the next three years (second, third, an d fourth) will be applied to the third year, and the sum of the subse quent three years (third, fourth, and fifth) will be applied to the fou rth year. This process will continue until the value of the final year is accounted for in the total.

(iii) The moving averages will be calculated by dividing each movi ng total by 3.

It is crucial to note that moving averages will not be calculated for the first and last year in the case of 3-yearly moving averages, and for the first two and last two years in t he case of 5-yearly moving averages.

Example:

From the following data taken Three-yearly moving average , calculate trend values :

Year	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022
Values	20	22	23	25	41	34	29	31	27	39	23	25	41

Solution:

Year	Values	Three Yearly Moving Totals	Three Yearly Moving Average (Trends or T)
2010	20	-	-
2011	22	65	21.66
2012	23	70	23.33
2013	25	89	29.67
2014	41	100	33.33
2015	34	104	36.67
2016	29	94	31.33
2017	31	87	29
2018	27	97	32.33
2019	39	89	29.67
2020	23	87	29
2021	25	89	29.67
2022	41	-	-

(2) Even Period Moving Averages-

The moving average is calculated after situating the moving totals i f it is to be calculated on the basis of an even period, such as 2, 4, o r 6 years. In the event that four-

year moving totals are to be computed, the subsequent methodolog y would be implemented: (i) Initially, four-year moving totals will be determined.

The initial sum will encompass the first four years, followed by the subsequent sum of the four years excluding the first year, and this procedure will be repeated.

The initial sum will be allocated between the second and third year s of the year, the second sum between the third and fourth years, an d so forth. (ii) Subsequently, these moving totals will be cantered. To achieve this objective, two-

period moving totals will be computed.

(iii) The moving totals of the two periods will be divided by 8.

Example:

From the following data taken Four-yearly moving average, calculate trend values:

Year	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022
Values	20	22	23	25	41	34	29	31	27	39	23	25	41

Solution:

Year	Values	Four Yearly Moving Totals	Two periods moving total centred	Moving Average (Trends or T)

2010	20	-	-	-
2011	22	-	-	-
		90	-	-
2012	23		200	25
		111		
2013	25		234	29.25
		123		
2014	41		252	31.5
		129		
2015	34		264	33
		135		
2016	29		256	32
		121		
2017	31		247	30.875
		126		
2018	27		246	30.75
		120		
2019	39		234	29.25
		114		
2020	23		242	30.25
		128	-	-
2021	25	-	-	-
2022	41	-	-	-

Advantages of the Moving Average Method:

(1) Simplicity: This method is straightforward to comprehend and i mplement, as it does not necessitate intricate mathematical calculat ions. (2) Objectiveness-

This method is objective in the sense that any individual who is wo rking on a problem with this method will obtain the same trend val ues. In this regard, it surpasses the free-hand curve method.

(3) Flexibility - This approach is sufficiently adaptable in comparis on to both semi-average and least squares.

It implies that the entire calculation will not need to be repeated if specific new values are incorporated.

Its sole consequence is the addition of additional trend values.

(4) Cyclic fluctuations are naturally eradicated if the period of mov ing averages coincides with the period of cyclical fluctuations in th e data. (5) Measurement of other components of time series— This method is employed to ascertain trend values, as well as seaso nal, cyclical, and irregular variations.

Limitations:

(1) The method is significantly limited by the inability to acquire tr end values for certain years at the beginning and conclusion. For example, the trend value for the first and last year is not obtain ed in a three-

yearly moving average, and the trend value for the first two and las t two years is not obtained in a four-yearly moving averages.

(2) Determination of the moving average period-

The determination of the moving average period is also a challengi ng endeavor under this method, particularly if the series lacks busi ness cycles.

(3) This method is not beneficial for forecasting and predicting val ues over time, as trend values are not articulated in terms of functio nal relationships. (4) Only applicable in standard variations In general, this approach is deemed suitable only for time series tha t exhibit consistent fluctuations.

In other words, this approach is not suitable for other circumstance s.

(5) The impact of extreme values-

Extreme values also have an impact on moving averages.
The method is superior to the free-hand curve and semi-average methods, despite these limitations.

Nevertheless, this approach is suitable for fitting a trend when (a) t he trend is linear, (b) the cyclical variations are consistent in both a mplitude and period, and (c) the purpose of the investigation does not entail current analysis or future forecasting.

10.6 METHOD OF LEAST SQUARES

This approach is regarded as one of the most effective methods for obtaining trend values.

The line of best fit for the time series is determined using algebraic equations under the assumption of least squares in this method.
This line may be represented as either a straight line or a parabolic curve.
The method of least squares is so named because the sum of square s of the deviations of the various locations of the trend line from th e original data is the least when compared to the sums of squares o f the deviations obtained by using any other line.

The method of least squares can be used to determine trend in three distinct ways:-

(1) Fitting a Parabolic or Non-linear Trend,

(2) Fitting a Straight Line Trend, and

(3) Semi-logarithmic or Exponential Curve.

(1) Fitting a Straight Line Trend –

For the purpose of fitting a straight line trend

The subsequent equation is employed in accordance with the least squares method:-

$Yc = a + bx$

Y = required trend value

X= unit of time

Where 'a' and 'b' are constants,

'a' is the difference between the point of origin (O) and the location where the trend line and Y-

axis intersect if the first year is considered the origin.
The arithmetic mean of the time series is denoted by 'a' if the middle year of the time series is considered the origin.
The slope of the trend line is denoted by the constant 'b'.
It is also referred to as the growth rate or decline rate, as it indicates the change in the trend line (Y) for each unit change in time (X).

The values of the constants 'a' and 'b' are determined using the following two equations:

$\Sigma Y = N a + b \Sigma X$

$\Sigma XY = a \Sigma X + b \boldsymbol{\Sigma X^2}$

The extended method of least squares is the term used to describe the utilization of equations in the aforementioned format.
Nevertheless, the value of ΣX is zero if deviations are derived precisely from the middle year of the time series. In this scenario, the aforementioned equations can be summarized as follows:-

a = ΣY/ N

b= $\boldsymbol{\Sigma XY/ \Sigma X^2}$

The brief method of least squares is the term used to describe the utilization of summarised versions of equations.

The brief method should be preferred if the query does not specify any contraindications and it is feasible to consider deviations preci sely from the middle year, as its calculation procedure is straightfo rward.

Procedure Least Squares Method-

The procedure for this technique is as follows:-

(i) A total of six columns are drawn, each of which represents a ye ar, value (Y), deviations (X), XY, $\boldsymbol{X^2}$, and trend values (Yc).
(ii) Initially, the column of X displays the time deviations for all ot her years from the exact middle or median year.
It is imperative to verify that the sum of the deviations is zero, or t hat ΣX = 0.

$\boldsymbol{\Sigma X^2}$ is calculated by squaring each deviation ($\boldsymbol{X^2}$) and summing th e resulting squares.

(iv) The multiplication (XY) of values (Y) and deviations (X) calc ulates ΣXY.

(v) The formula for calculating the value of the constant 'a' involve s ΣY / N.

The symbol N is used to represent the number of years (time units)

(vi) The formula for determining the value of 'b' is as follows:
ΣXY / ΣX

(vii) The formula a + bX is applied to each year to derive trend val ues (Yc).

It is imperative to verify that the sum of the original values (ΣY) is equivalent to the sum of the trend values (ΣYc, or ΣY).

Example:

From the following data trend values.

Year	2018	2019	2020	2021	2022
Production ('000 Qntls.)	31	39	45	55	61

Solution:

Year	Production ('000 Qntls.)	Deviation from 2020	Squares	Product of X and Y	Trend Value
	Y	X	X^2	XY	Yc = a+ bX
2018	31	-2	4	-62	46.2 + 7.6 x -2 = 31
2019	39	-1	1	-39	46.2 + 7.6 x -1 = 38.6
2020	45	0	0	0	46.2 + 7.6 x 0 = 46.2
2021	55	1	1	55	46.2 + 7.6 x 1 = 53.8
2022	61	2	4	122	46.2 + 7.6 x 2 = 61.4
N= 5	∑Y= 231		ΣX^2= 10	∑XY= 76	

a = ΣY/ N = 231/5 = 46.2

b= *$\Sigma XY/ \Sigma X^2$ = 76/5 = 7.6*

Example:

From the following data trend values.

Year	2017	2018	2019	2020	2021	2022
Sales in Lakhs	131	139	145	155	161	180

Year	Sales in Lakhs	Deviation from 2019.5 and multiplied by 2	Squares	Product of X and Y	Trend Value
	Y	X	X2	XY	Yc = a+ bX
2017	131	-5	25	-655	151.83 + 4.585 x -5 = 128.91
2018	139	-3	9	-417	151.83 + 4.585 x -3 = 138.08
2019	145	-1	1	-145	151.83 + 4.585 x -1 = 147.25
2020	155	1	1	155	151.83 + 4.585 x 1 = 156.42
2021	161	3	9	483	151.83 + 4.585 x 3 = 165.59
2022	180	5	25	900	151.83 + 4.585 x 5 = 174.76
N= 6	∑Y= 911		∑X2= 70	∑XY= 321	

a = ΣY/ N = 911/6 = 151.83

b= *ΣXY/ ΣX^2 = 321/70 = 4.585*

Benefits of the Least Squares Method:

(1) Completely objective: This method is entirely objective, as it calculates trend values using well-defined mathematical principles

and formulae. This procedure is devoid of any potential for personal bias.

(2) Forecasting-The least squares method's equation of a straight line establishes a functional relationship between the x and y series, which enables the prediction of future values.

(3) The moving average method is unable to determine the trend values for the entire period, whereas the least squares method provides the trend values for the entire time period.

(4) Line of best fit - The trend line derived by this method is the line of best fit, as the sum of positive and negative deviations of the original data from this line is zero and the sum of squares of deviations is the minimal number.

(5) Change rate calculation: This method can be used to determine the annual growth rate or decline rate if the data is collected on a yearly basis.

Limitations of the Least Squares Method:

(1) Complicated and tiresome-This method is complex and laborious in terms of mathematical calculations.

(2) Inflexibility - The equation of trend is frequently altered, and the computations must be repeated if even a single value is added or removed from the series.

(3) Error in equation selection-Errors may result from the improper selection of a trend equation (e.g., linear, parabolic, or any other type).

(4) Prediction limitations: This method relies on long-term trends and disregards the influence of seasonal, cyclical, or irregular fluctuations.

10.7 SHORT-TERM FLUCTUATION MEASUREMENT

The values of a time series are the sum of the short-term fluctuations and the long-term trend. Thus, by subtracting trend values from the original values, short-term fluctuations can be determined if it is presumed that there is no irregular or arbitrary fluctuation in the data. The trend values are determined through either the least squares method or the moving average method.

Example:

From the following data taken Five-yearly moving average, calculate trend values and short term fluctuations:

Year	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022
Values	20	22	23	25	41	34	29	31	27	39	23	25	41

Solution:

Year	Values (O)	Five Yearly Moving Totals	Five Yearly Moving Average	Short Term Fluctuations (O-T)

			(Trends or T)	
2010	20	-	-	-
2011	22	-	-	-
2012	23	131	26.2	-3.2
2013	25	145	29	-4
2014	41	152	30.4	10.6
2015	34	160	32	2
2016	29	162	32.4	-3.4
2017	31	160	32	-1
2018	27	149	29.8	-2.8
2019	39	145	29	10
2020	23	155	31	-8
2021	25	-	-	-
2022	41	-	-	-

10.8 APPLICATIONS OF TIME SERIES ANALYSIS IN BUSINESS.

In the business world, time series analysis is a potent instrument for predicting future trends and making data-driven decisions. The following are a few critical applications:

Sales Forecasting:

Forecast future sales by employing historical data.

Efficiently plan inventory and oversee supply chain operations.

Financial Analysis:

Forecast economic indicators, exchange rates, and stock prices.

Examine revenue, expense, and profit trends to inform investment and budgeting decisions.

Planning for Demand:

Anticipate the demand for products and services among customers.

Optimize inventory levels and minimize holding costs.

Marketing analytics:

Evaluate the effectiveness of marketing campaigns over time.

Based on customer behavior patterns, optimize marketing strategies.

Risk Management:

Foresee prospective market fluctuations and financial hazards.

Formulate strategies to reduce the hazards associated with investment portfolios.

Operations Management:

Evaluate and forecast operational metrics, including equipment maintenance requirements and production rates.

Enhance the efficacy of processes and optimize resource allocation.

Customer Behavior Analysis:

Monitor and anticipate the purchasing habits and preferences of customers.

Enhance the targeting and segmentation of customers.

Energy Consumption Forecasting:

Optimize energy procurement and anticipate future energy requirements.

Formulate strategies for cost reduction and energy conservation.

Healthcare Management:

Predict the rates of patient admission and the necessary resources.

Enhance the scheduling and personnel of hospitals.

Management of the Supply Chain:

Forecast the demand for basic materials and oversee supplier relationships.

Reduce transportation expenses and optimize logistics.

Businesses can enhance overall efficiency and profitability, improve decision-making processes, and acquire valuable insights by utilizing time series analysis.

10.8 LET US SUM UP :

Time series analysis is the examination of data points that have been gathered or recorded at regular time periods. It is a potent instrument in the corporate realm for predicting, tracking patterns, and making well-informed choices. Below are many essential techniques and their respective uses:

Time series analysis encompasses several methods, one of which is Moving Averages (MA).

The Simple Moving Average (SMA) is a mathematical technique that reduces the impact of short-term variations in order to uncover long-term trends. This tool is valuable for making short-term predictions and seeing patterns in sales data.

An Exponential Moving Average (EMA) assigns more significance to recent data points. This method is very efficient in catching the most up-to-date patterns and may be used for analyzing stock prices.

Autoregressive Integrated Moving Average (ARIMA)

The ARIMA model is a statistical model that combines autoregression, differencing, and moving averages. Applicable to dynamic time series data and often used for economic predictions, such as GDP or inflation rates.

Periodic ARIMA (SARIMA) is an extension of the ARIMA model that takes into consideration the presence of seasonally. Valuable for predicting retail sales in situations where seasonal trends play a major role.

Exponential smoothing

Simple Exponential Smoothing is a technique that uses a factor to smooth out previous data in order to predict future values. Well-suited for time series data that lacks a trend or seasonal rhythm, such as daily website traffic.

The Holt's Linear Trend Model is an extension of exponential smoothing that is designed to capture linear trends. Beneficial for making predictions over extended periods, such as determining inventory levels or projecting product demand.

The Holt-Winters Seasonal Model is a forecasting method that takes into account both the trend and seasonality of a time series. Seasonal items or services, such as garment sales, may be accurately forecasted using this method.

Time series decomposition using the Seasonal Decomposition of Time Series (STL) method.

STL Decomposition is a method that separates a time series into its seasonal, trend, and residual parts. This is beneficial for comprehending fundamental trends and enhancing predictions in sectors such as tourism or energy.

Vector Autoregression (VAR)

A VAR model is a statistical model that examines the relationships between various time series variables that have an impact on each other. Valuable in the field of econometrics for comprehending the connections between economic indices, such as unemployment and inflation.

GARCH (Generalized Autoregressive Conditional He t eroske dasti city) is a statistical model used to analyze and predict the volatility of a time series data, taking into account the conditional he ter oske dasti city.

The GARCH model is used to capture the phenomenon of volatility clustering in time series data. Utilized in financial markets to predict stock market volatility and mitigate risk.

Business Applications: Sales Forecasting

Utilizing ARIMA, moving averages, or exponential smoothing techniques to forecast forthcoming sales and effectively handle inventory management.

Financial analysis

Utilizing GARCH and VAR models to comprehend stock price volatility, manage portfolios, and evaluate financial risks.

Forecasting and estimating future demand for products or services.

Utilizing seasonal decomposition and Holt-Winters models for predicting product demand and enhancing supply chain operations.
Analysis of customer behaviour.

Examining temporal data of client engagements to detect buying trends and enhance advertising tactics.

Efficiency in operations
Analyzing time series data from operational processes to detect patterns, identify obstacles, and pinpoint areas for improvement.

Economic forecasting
Employing ARIMA and VAR models for predicting economic indicators, such as GDP growth or inflation rates, to inform strategic business choices.

10.9 KEY WORDS:

Simple Moving Average (SMA) : is a mathematical technique that reduces the impact of short-term variations in order to uncover long-term trends. Valuable for making predictions in the near future and seeing patterns in sales data.
Exponential Moving Average (EMA): assigns more significance to recent data points. This method is very efficient in catching the most up-to-date patterns and may be used for analyzing stock prices.

10.10 ANSWERS TO CHECK YOUR PROGRESS :

1. The additive model may be expressed as Y(t) =……., where Y(t) represents the dependent variable at time t.

2. An ……..model is often used when the seasonal variation remains consistent across the series.

3. In an additive model, as the trend component grows, the seasonal variations stay……..

4. The multiplicative model may be expressed as Y(t) =……., where Y(t) represents the dependent variable, Trend represents the long-term trend, Seasonal represents the seasonal component, and Irregular represents the random fluctuations.

5. The …….component pertains to the extended oscillations or waves included in the data series.

Answer:

1. Trend + Seasonal + Irregular
2. Additive
3. Constant
4. Trend × Seasonal × Irregular
5. cyclical

10.11 TERMINAL QUESTIONS:

Q1. Explain Additive and Multiplicative model.
Q2. Discuss Method of least squares.
Q3. Distinguish between Additive and Multiplicative model.
Q4. Explain Moving average method.

BLOCK IV: PROBABILITY

UNIT 11: CONCEPT OF PROBABILITY AND ITS USES IN BUSINESS DECISION-MAKING

Structure

11.0 Objectives

11.1 Introduction

11.2 Important Terminology in Probability

11.3 Types of Events

11.4 Types of Probability

11.5 Probability Expression

11.6 Uses of Probability in Business Decision-Making

11.7 Let Us Sum Up

11.8 Key Words

11.9 Answers to Check Your Progress

11.10 Terminal Questions

11.0 OBJECTIVES

After studying this unit, you should be able to:

• Understand the notion of probability.

• Elucidate the various types of occurrences and their corresponding probability.

• Gain information on how to express probability.

• Examine the practical uses of probability in the process of making informed business choices.

11.1 INTRODUCTION:

Probability refers to the likelihood or chance of an event occurring. Probability is a mathematical discipline that focuses on the occurrence of unpredictable events. The value is represented on a scale ranging from zero to one. Probability is a mathematical concept used to forecast the likelihood of occurrences occurring. Probability refers to the degree of likelihood that an event will occur. This is the fundamental concept of probability theory, which is also used in probability distribution. In probability distribution, you will study the likelihood of different outcomes given a random experiment. In order to determine the likelihood of a singular event occurring, it is necessary to ascertain the total number of potential possibilities.

Probability is a quantitative assessment of the chances or probability of an event happening. Several occurrences are inherently unpredictable. Using it, we can only forecast the probability of an event occurring, indicating its likelihood. The probability of an occurrence may be expressed as a value between 0 and 1. A probability of 0 signifies that the event is impossible, while a probability of 1 implies that the event is certain to occur. Probability is a crucial subject for Class 10 students since it covers fundamental principles related to this discipline. The cumulative probability of all occurrences inside a given sample region equals 1.

When we toss a coin, there are only two potential outcomes: either we obtain a Head or a Tail (H, T). However, when two coins are thrown, there will be a total of four potential outcomes, namely {(H, H), (H, T), (T, H), (T, T)}.

The probability formula is described as the likelihood of an event occurring, which is determined by dividing the number of favorable outcomes by the total number of possible possibilities.

11.2 IMPORTANT TERMINOLOGY IN PROBABILITY

Experiment: An experiment is a systematic procedure conducted to produce a certain outcome or result.

Sample space refers to the set of all possible outcomes that might occur during an experiment. For instance, in the context of coin tossing, the sample space consists of two possible outcomes: heads and tails.

Favorable Outcome: A favorable outcome refers to an occurrence that has produced the expected or projected consequence. When two dice are rolled, the favorable results for obtaining a total of 4 are (1,3), (2,2), and (3,2). The coordinates are (3,1).

Random Experiment: A random experiment refers to an experiment that has a clearly specified collection of possible results. When we flip a coin, we can determine whether it will land on heads or tails, but we cannot predict which outcome will occur.

An event in a random experiment represents the entire number of possible outcomes.

Equally probable occurrences refer to events that have an equal chance or likelihood of happening. The result of one particular event does not have any influence on the outcome of another event. When we flip a coin, we have an equal likelihood of obtaining either a head or a tail.

An exhaustive event occurs when the set of all potential outcomes of an experiment is equal to the sample space.

11.3 TYPES OF EVENTS

In the field of Probability, we encounter several categories of occurrences. An event refers to the set of possible events that might occur as a result of an experiment. Probability is applicable in several disciplines. Probability pertains to the occurrence of an unpredictable event. The probability of an event E, denoted as P(E), is calculated by dividing the number of favorable outcomes of E by the total number of potential outcomes. The concept of probability has significant importance in the JEE examination. This article aims to provide a comprehensive understanding of different sorts of occurrences in probability via the use of illustrative examples.

Confirmed occurrence

It is an inevitable occurrence that takes place whenever an experiment is undertaken. For instance, the occurrence of obtaining a tail when a coin is flipped. The probability of an event that is certain to occur is 1.

Illustration:

The probability of an occurrence that encompasses all possible outcomes of an experiment, also known as the sample space, is equal to 1.

Unattainable occurrence

An event with a probability of 0 is classified as an impossible event.

Example: The occurrence of obtaining a 7 when a die is rolled is seen to be impossible. The reason behind this is because the possible results of rolling a die are {1, 2, 3, 4, 5, 6}.

Autonomous occurrence

Occurrences that are not affected by the result of a previous event are referred to be independent occurrences.

The occurrence of obtaining a tail when flipping one coin and the occurrence of obtaining a head when flipping another coin.

Conditional occurrence

Events in which the result of the first event affects the result of the second event are referred to as dependent events.

Example: If we choose two colored marbles from a bag without replacing the first marble before drawing the second marble, then the result of the second draw will be influenced by the result of the first draw.

Exclusive Event

These occurrences are mutually exclusive and cannot occur simultaneously. Simultaneous occurrence is not possible.

The occurrences of obtaining either heads or tails are mutually exclusive when flipping a coin.

Complementary event refers to an event that occurs when the outcome of another event does not occur.

For each given event A, another event A', represents the remaining components of the sample space S. A' is equal to S minus A.

Suppose we have a sample space, S, which consists of the first 10 natural numbers: S = {1, 2, 3, 4, 5, 6, 7, 8, 9, 10}. Let A be the occurrence of selecting a number from S that is both even and smaller than 10. The set A is defined as {2, 4, 6, 8}.

Therefore, the set A' is equal to the set S minus the set A, resulting in the set {1, 3, 5, 7, 9, 10}.

Composite Occurrence

A compound event refers to an occurrence that consists of many sample points.

Given the set S = {1, 2, 3, 4, 5, 6}, we get two subsets: E1 = {1, 3, 6} and E2 = {2, 6}. Therefore, E1 and E2 indicate compound events.

Comprehensive Occurrence

The events E1, E2,……., En are considered exclusive if their union, E1 ∪ E2 ∪…….∪ En, is equal to the sample space S.

Suppose we have event E1, which represents the outcome of receiving an even number when tossing a dice, and event E2, which represents the outcome of obtaining an odd number.

Let E1 be the set containing the elements {2, 4, 6}, and let E2 be the set containing the elements {1, 3, 5}.

The union of sets E1 and E2, denoted as E1 ∪ E2, is equal to the set {1, 2, 3, 4, 5, 6}, which represents the sample space S.

Therefore, E1 and E2 are occurrences that cover all possible outcomes.

Basic occurrence

A simple event in probability refers to an event that consists of just one outcome from the sample space.

Given the sets S = {1, 2, 3, 4} and E = {3}, it may be concluded that E is a simple event.

11.4 TYPES OF PROBABILITY

Classical probability, commonly referred to as theoretical probability, is a field within probability theory that focuses on scenarios where all potential outcomes of an event have an equal

likelihood of occurring. This kind of probability is often used in situations when the sample space is limited and each possible result has an equal likelihood of occurring.

The classical probability formula is expressed as follows: P(A) = Number of favorable outcomes for event A divided by the total number of potential outcomes. The user's text is simply "A".

The probability of occurrence A is denoted as P(A).

The numerator denotes the quantity of outcomes that are advantageous to event A.

The denominator corresponds to the cardinality of the sample space, which reflects the total number of potential outcomes.

An exemplary instance is a symmetrical six-sided dice. The likelihood of obtaining any particular number (1, 2, 3, 4, 5, or 6) when rolling a die is 1/6, as there are six equally probable options.

Essential aspects of classical probability:

Equiprobable Outcomes: Classical probability postulates that every event inside the sample space has an equal chance of occurring. This assumption may not be valid in scenarios where some outcomes have a higher probability than others.

Classical probability is often used when the sample space is limited to a certain number of outcomes. For instance, a pack of playing cards, a hexahedral die, or the act of tossing an unbiased coin.

Probabilities are constrained to be non-negative, meaning they cannot have negative values, and they must also total up to 1. This implies that the total of the odds of every conceivable result inside the sample area is equal to 1.

Classical probability is applicable in straightforward and well defined circumstances, but it may not be appropriate for intricate

scenarios where outcomes are not equally probable or when there are an unlimited number of potential possibilities. In such instances, other disciplines of probability theory, such as empirical or subjective probability, may be more suitable.

Axiomatic Probability:

Axiomatic Probability refers to a method of quantifying the likelihood of an event occurring. In this strategy, probabilities are assigned after predefining certain axioms. The purpose of this is to discretize the event, so facilitating the computation of its occurrence or non-occurrence.

Axiomatic Probability

Probability is a set function $P(E)$ that assigns to every event E a number called the "probability of E" such that:

1. The probability of an event is greater than or equal to zero

$$P(E) \geq 0$$

2. The probability of the sample space is one

$$P(\Omega) = 1$$

Subjective probability

Subjective probability is a kind of probability that produces outcomes depending on an individual's own opinion and viewpoint of a certain result or event. Subjective probability relies on individual experiences and beliefs rather than quantitative calculations or historical evidence. It encompasses a significant amount of subjective prejudice and is prone to vary across individuals.

Subjective probability is used by a company when it encounters uncertain factors or lacks enough knowledge to determine the likelihood of an event happening. In some cases, certain sectors, such as marketing and economics, consider it to be the only option.

The advantage of subjective probability lies in its little reliance on quantitative and historical facts. It is useful for making estimates or forecasts and draws upon important human experience and expert opinion.

Empirical probability

An empirical probability, also known as experimental probability, is strongly correlated with the relative frequency of an occurrence. Empirical probability relies on the frequency of a certain result in a sample set to estimate the likelihood of that outcome happening again. The frequency of "event X" occurring in 100 trials will determine the chance of event X happening.

The empirical probability formula calculates the ratio of the number of times a desired event occurs to the total number of attempts made to achieve it. An instance of this would be when I rolled the dice on three separate occasions, and each time I obtained a sum of 12. This resulted in a statistical probability of 12/12, which equates to 100%. This calculation illustrates the limitation of empirical probability.

Conditional probability

Conditional probability is the chance of an event or result happening, given that a preceding event or outcome has already occurred. Conditional probability is determined by multiplying the likelihood of the previous event with the revised probability of the subsequent occurrence, which is dependent on the preceding event.

Conditional probability may be distinguished from unconditional probability. Unconditional probability is the chance of an event occurring without any consideration of other occurrences or circumstances.

Conditional probabilities are dependent on the occurrence of a prior outcome or event. Conditional probability examines the

interrelationship between such occurrences. Conditional probability refers to the chance of an event or result happening, given that another event or preceding outcome has already occurred.

Two occurrences are considered independent if the occurrence of one event does not influence the likelihood of the other event happening. If the occurrence or non-occurrence of one event has an impact on the likelihood of the other event happening, then the two occurrences are considered dependent. If events are independent, the occurrence of event B is not influenced by the outcome of event A. A conditional probability pertains to occurrences that are interdependent.

Conditional probability is often represented as the "probability of event A occurring given that event B has occurred," denoted as P(A|B).

P(B|A) = P(A∩B) / P(A) or P(B|A) = P(A and B) / P(A)

11.5 PROBABILITY EXPRESSION

Probability is a mathematical quantity that can be represented in a variety of ways that are equivalent in terms of their mathematical value. These forms may be expressed as ratios, fractions, or percentages. For instance, the likelihood of receiving a head in a coin strike can be expressed as 1/2, 0.5, 50%, or 1:1.

It is also important to mention that the probability of an event occu rring is always between 0 and 1, meaning that $0 < p < 1$.
The event is considered impossible if the probability is zero.
Conversely, if the probability is 1, it is not considered a probability ; rather, it is referred to as a certainty.

Example:

(i) What is the chance of drawing a King in a draw from a pack of 52 cards?

(ii) A ball is drawn from a bag containing 6 red, 4 white and 5 blue balls. Determine the probability that it is (a) red (b) white (c) black.

Solution:

Solution:

(i) Total cards in a pack or total possible events = 52

Cards of king or favourable events = 4

Probability of drawing a king (P) = 4/52 = 1/13

(ii) Total events = 6+4+5=15 balls

If red, white and black balls are denoted as R, W and B respectively, Then:

P(R)= 6/15 = 2/5

P(W)= 4/15

P(B) = 5/15 = 1/3

11.6 USES OF PROBABILITY IN BUSINESS DECISION-MAKING

Probability is essential in corporate decision-making since it offers a structured approach to assess and measure uncertainty. Probability is used in several ways in commercial decision-making:

Evaluation of potential hazards:

Probability enables organizations to evaluate the probability of different risks and uncertainties. Through the allocation of probability to various situations, firms may make well-informed choices about the management and reduction of risks.

Prediction:

Probability plays a crucial role in predicting future occurrences or outcomes. Businesses may use probability models to evaluate the probability of various events and make informed decisions on sales, market trends, or project timetables.

Analysis of decisions:

Probabilities are allocated to distinct potential outcomes in decision analysis, and various methods such as decision trees or other models are used to assess alternative courses of action. This facilitates the selection of the most logical option by considering anticipated values.

Financial Planning:

Probability is essential in the process of financial modeling and planning. Probability distributions are used by businesses to model various financial situations, evaluate possible returns, and make well-informed choices about investments, budgeting, and resource allocation.

Marketing and Consumer Behavior:

Probability models are used in marketing to forecast client behavior, reaction rates, and the efficacy of marketing initiatives. This allows firms to efficiently deploy resources and customize their tactics to optimize profits.

Quality Assurance:

In the realm of manufacturing and production, probability is used to evaluate the possibility of flaws or mistakes occurring in a product. Statistical methodologies, such as control charts, depend on probability to ascertain acceptable levels of quality and maintain uniformity.

Logistics and operations management:

Probability is used in supply chain management to evaluate the probability of interruptions, delays, and fluctuations in the supply chain. This data assists firms in optimizing inventory levels, strategizing for unforeseen circumstances, and maintaining streamlined operations.

HR:

HR use probability models to forecast employee attrition, evaluate the effectiveness of recruiting methods, and make well-informed choices about workforce planning and talent management.

Insurance and risk management:

Insurance firms significantly depend on probability to evaluate risks and establish premium rates. Probability models assist in calculating the probability of certain occurrences, such as accidents or natural disasters, and determining the appropriate amount of coverage.

Management of projects:

Probability is used in project management to evaluate the probability of fulfilling project deadlines, adhering to budget constraints, and attaining project objectives. This facilitates improved project planning and risk reduction during the whole duration of the project.

To put it simply, probability offers a methodical and quantitative strategy for addressing uncertainty in business. It enables firms to enhance decision-making, mitigate risks, and optimize resource allocation.

11.7 LET US SUM UP:

Probability is a mathematical discipline that focuses on the probability of certain events happening. Uncertainty is measured and represented as a numerical value ranging from 0 to 1, with 0 indicating that an event will not occur and 1 indicating that it will certainly occur. As the probability increases, the likelihood of the event happening also increases.

Essential Principles in Probability
Probability is determined by dividing the number of favorable events by the total number of potential possibilities. The chance of rolling a 3 on a fair six-sided die is 1/6.

Conditional probability is a statistical metric that quantifies the likelihood of an event happening, taking into account the occurrence of another event. For example, if you are aware that it is raining, the likelihood of someone possessing an umbrella is greater compared to when you have no knowledge of the weather.

Independent and Dependent Events: Independent events are not influenced by each other's probability (e.g., the outcome of rolling a die does not affect the outcome of flipping a coin). Dependent events are interrelated and affected by one another, for as when pulling cards from a deck without replacing them.

Probability distributions represent the way in which probabilities are spread out across different potential values. Typical distributions are the normal distribution (also known as the bell curve), binomial distribution, and Poisson distribution.

Application in Business Decision-Making Risk Mitigation: Probability enables organizations to evaluate the probability of different hazards and their possible consequences. Insurance firms use probability to calculate insurance rates and coverage choices by assessing the likelihood of claims.

Market research use probability to evaluate and assess customer behavior and preferences. Businesses may use probability to forecast sales trends, analyze client buying habits, and evaluate the efficacy of marketing programs.

Financial Forecasting: Companies use probability models to predict financial performance, including the estimation of future revenues, expenses, and profits. This contributes to the process of allocating financial resources and developing long-term plans.

Quality Control: In the field of manufacturing, probability is used in quality control procedures to assess the probability of faults and guarantee that goods adhere to established quality criteria.

Decision analysis utilizes probability to facilitate decision-making in situations where there is ambiguity. It involves assessing several scenarios and their respective probabilities. Methods such as decision trees and Monte Carlo simulations are used to assess possible outcomes and make well-informed choices.

Resource allocation in businesses involves using probability to maximize the distribution of resources. This includes making decisions on how to spend funds or deploy staff based on the anticipated return on investment or the success rates of projects.

Probability is a fundamental tool for making well-informed decisions. It allows us to measure uncertainty and evaluate the potential risks and rewards of various choices.

11.8 KEY WORDS:

Probability: refers to the likelihood or chance of an event occurring.
Conditional probabilities: are dependent on the occurrence of a prior outcome or event.

11.9 ANSWERS TO CHECK YOUR PROGRESS:

1. The …….of an occurrence is determined by dividing the number of positive outcomes by the total number of outcomes.
Answer: probability
2. The likelihood of obtaining heads while flipping a fair coin is……
Answer: 1/2
3. The likelihood of rolling a 3 while using a fair six-sided die is……..
Answer: 1/6
4. The likelihood of getting a king from a regular deck of 52 cards is……...
Answer: 1/13
5. The chance of selecting a vowel from the English alphabet, assuming that each letter has an equal likelihood of being chosen, is……...
Answer: 5/26

11.10 TERMINAL QUESTIONS:

Q.1.Explain Concept of probability?

Q2. Explain term random experiment and sample space?

Q3. Explain Modern Approach of probability.

Q.4. Explain Empirical or statistical approach of probability.

UNIT 12: ADDITION, MULTIPLICATION THEOREM OF PROBABILITY AND BINOMIAL THEOREM

Structure

12.0 Objectives

12.1 Introduction

12.2 Addition Theorem of Probability

12.3 Applications of Addition theorem

12.4 Multiplication Theorem

12.5 Applications of Multiplication theorem:

12.6 Binomial Theorem:

12.7 Applications of Binomial Theorem:

12.8 Let Us Sum Up

12.9 Key Words

12.10 Answers to Check Your Progress

12.11 Terminal Questions

12.0 OBJECTIVES

After studying this unit, you should be able to:

- Comprehend the Addition theorem of probability
- Comprehend the Multiplication theorem of probability
- Elucidate the applications of the Addition and Multiplication theorems of probability
- Elaborate on the notion and applications of the Binomial Theorem

12.1 INTRODUCTION

The probability of the occurrence of at least one of two events is determined using the addition theorem of probability.

The probability of the joint occurrence of two events is determined using the multiplication theorem of probability.

Binomial theorems offer fundamental tools for addressing a diverse array of problems in probability, algebra, and beyond for any two events, A and B.

12.2 ADDITION THEOREM OF PROBABILITY

Probability theory employs a statistical feature to explain the occurrence of one or more events during a single activity. Consider a random experiment involving the selection of a card from a deck. The event of interest is the selection of a card that is either a jack or a queen. It is not feasible to draw both the jack and queen in the same draw. The result will be either a jack or a queen. The addition rule is used to address this particular probability issue.

The addition rule is used to determine the combined probability of one or more occurrences in a situation that presents an either-or statement structure.

Put simply, the addition theorem calculates the probability of at least one of the specified occurrences occurring. The notion of union is used to calculate the likelihood of at least one event occurring. The idea of union is fundamental in mathematical probability theory and set theory. The depiction of this union resembles the letter U in the alphabet.

When two events cannot happen at the same time inside a task and cannot have the same result, they are referred to as mutually exclusive events. If many events have the same conclusion, then

those events cannot be considered mutually exclusive. It may be described as a non-mutually exclusive occurrence.

The Addition Theorem of Probability states that given mutually exclusive events, the probability of the union of these events is equal to the sum of their individual probabilities. Statement: If A and B are two mutually exclusive occurrences, then the likelihood of either A or B occurring is equal to the sum of the probabilities of A and B individually.

This may be expressed as:
The probability of the union of events A and B, denoted as P(A∪B), is equal to the probability of event A or event B, which may be calculated by adding the individual probabilities of A and B, denoted.

P(A∪B)=P(A or B)=P(A)+P(B)

Example 1: A single card is selected at random from a standard deck of 52 cards. Determine the chance of selecting a card that is either a diamond or an ace of clubs.

Solution 1: Let A: Event of drawing a card of diamond.

Let B: Event of drawing an ace of club.

The probability of drawing a card of diamond P(A) = 13/52.

The probability of drawing an ace of club P(B) = 1/52.

Since the events are mutually exclusive, probability of drawing a card being a diamond or an ace of club is:

$$P(A \cup B) = P(A) + P(B) = \frac{13}{52} + \frac{1}{52} = \frac{14}{52} = \frac{7}{26}$$

Additional Probability Theorem for Events that are not mutually exclusive.

The aforementioned probability theorem does not apply to occurrences that are not mutually exclusive.

Now, let's get into the addition theorem of probability for occurrences that are not mutually exclusive.

The Addition Theorem of Probability applies to events that are not mutually exclusive. Statement: If A and B are non-mutually exclusive occurrences, the probability of either A or B or both occurring is equal to the sum of the probabilities of A and B, minus the likelihood of the events that are common to both A and B.

The given information may be expressed as:
The probability of the union of events A and B is equal to the sum of the probabilities of A and B, minus the probability of their intersection.

$P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Example 2: A card is selected randomly from a deck consisting of 52 cards. Determine the likelihood of selecting a card that is either a club or a queen.

Solution 2: Let A: Event of drawing a card of club.

Let B: Event of drawing a queen card.

The probability of drawing a card of club P(A) = 13/52.

The probability of drawing a queen card P(B) = 4/52.

As, one of the queen is club, so, the events are not mutually exclusive.

So, the probability of drawing queen of club = $P(A \cap B) = 1/52$

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

$$= \frac{13}{52} + \frac{4}{52} - \frac{1}{52}$$

$$= \frac{16}{52} = \frac{4}{13}$$

12.3 APPLICATIONS OF ADDITION THEOREM

The addition theorem has a wide range of applications in a variety of disciplines, particularly in the context of probability and trigonometry. The following applications are noteworthy:

1. Probability Theory: The addition theorem in probability is a tool that facilitates the calculation of the probability of the union of two events. The addition theorem is as follows for two events, A A and B B:

$P(A \cup B) = P(A) + P(B) - P(A \cap B)$

P(A∪B) = P(A) + P(B) - P(A∩B)

Utilizations:

Risk Management: Determine the likelihood of the occurrence of multiple risks in the financial and insurance sectors.

Quality Control: Evaluate the probability of defects in manufacturing processes.

Market Research: Evaluate the likelihood that customers will favor at least one of several products.

2. Trigonometry: The trigonometric values of the sum or difference of angles are determined using the addition theorems for trigonometric functions (e.g., sine and cosine). For instance,

$\sin(A+B) = \sin(A)\cos(B)+\cos(B)\sin(B)$ sin(A+B)=sinAcosB+cosAsinB

cos(A+B)=cosA cosB−sinA sinB

Utilizations:
Engineering: The development and examination of oscillatory systems, including mechanical vibrations, sound waves, and AC circuits.

Physicists: Address wave interference issues and the resolution of vector components in mechanics.

Signal Processing: The process of analyzing and synthesizing signals in communication.

3. Quantum Mechanics: The addition theorem for spherical harmonics is employed in the investigation of angular momentum in quantum mechanics. Combining angular momenta from various sources is advantageous.

Utilizations:
Atomic and Molecular Physics: Determine the probability distributions of electrons in the vicinity of nuclei.
Nuclear Physics: Investigate the properties and interactions of nuclear particles.

4. Statistics:
The addition formula for expected values and variances in statistics can be employed to combine the expectations and variances of random variables.

Predictive modeling involves the integration of predictions from multiple models.

Risk Analysis: Combine risk metrics from various sources.

5. Computer Science: In computer graphics, the addition theorems for trigonometric functions are employed in rotation matrices and transformations.

Applications: 3D Modeling and Animation: Rotate objects in 3D space.

Game Development: Develop character movements that are realistic in terms of physics.

6. Signal Processing: Fourier analysis employs the addition theorems to decompose signals into their constituent frequencies.
Audio Processing: The process of analyzing and manipulating sound signals.

Image Processing: The process of filtering and transforming images.
Consequently, the addition theorem is essential for the resolution of real-world problems and the simplification of complex calculations in a variety of fields.

12.4 MULTIPLICATION THEOREM

Probability theory employs a statistical attribute to determine the likelihood of occurrences occurring in many tasks. Consider the scenario where two dice are rolled consecutively. We need to determine the probability of obtaining a sum of 3 on the first die and rolling an odd number on the second dice. Two tasks are occurring simultaneously: the first task involves rolling a dice for the first time, and the second task involves rolling a die for the second time. We need to determine the chance of both occurrences occurring. These problems may be resolved by using the multiplication probability theorem.

The multiplication probability theorem is used to determine the compound probability of two or more occurrences involving many tasks.

The Multiplication chance Theorem is used to determine the chance of occurrences happening in various tasks. The idea of intersection is used to determine the likelihood of occurrences that occur simultaneously. The idea of intersection is fundamental in mathematical probability and set theory, much as the concept of union. The symbol ∩ represents it.

Two events may occur concurrently in two separate activities. If the occurrence of an event is influenced by the outcome of another trial, it is considered a dependent event. Otherwise, it is classified as an independent event. The mathematical formulation for independent and dependent variables will exhibit modest variations. The terms used to describe these concepts are the multiplication rule for dependent events and the multiplication rule for independent events, respectively.

The multiplication theorem of probability asserts that when two occurrences are dependent and must occur concurrently, the likelihood of both events occurring is the product of their individual probabilities. The conditional probability of the simultaneous occurrence of events A and B, assuming that event A has already occurred, may be calculated by multiplying the individual probabilities of each event, as stated by the probability multiplication theorem.

The multiplication rule may be used to assess the probability of occurrences A and B occurring simultaneously when they happen independently.

Now, let's explore the formulation of multiplication probability for both dependent and independent events.

Dependent events are characterized by the interdependence of one event's occurrence or non-occurrence on the result of another event. The previously mentioned multiplicative theorem is not applicable in these instances.

When such events happen, the likelihood is referred to as the conditional probability and may be calculated using
P(A/B)=P(AB)/P(B)) or P(A∩B)/P(B)
The conditional probability of event B given event A is likewise.
P(B/A)=P(A∩B)/P(A)
Example 1: What is the probability of getting the same outcome on two consecutive rolls of a dice?

Solution: A is the probability of any side

B is the probability of landing on the same side once more.

The formula for Independent Success

$$P(A \cap B) = P(A) \times P(B)$$

$$P(A \cap B) = \frac{1}{6} \times \frac{1}{6}$$

$$P(A \cap B) = 0.027 \times 100$$

$$P(A \cap B) = 2.7$$

The Multiplication Theorem of Probability Events that are neither influenced nor affected by each other. The theorem states that the chance of two independent events occurring simultaneously is equal to the sum of the probabilities of each individual event.
P(A and B)=P(A)×P(B)
P(AB)=P(A)×P(B)
It is also feasible to extend the theorem to three or more independent occurrences.
P(A∩B∩C)=P(A)×P(B)×P(C)
Example 2: There are a total of 15 black balls and 10 white balls in the box. Consecutively, two balls are extracted from the box without

any alteration between games. What is the likelihood that both of the balls drawn will be white?

Solution: Let P and Q represent the first and second draws, respectively, in which a white ball is drawn from the box without replacement.

White ball in first draw $P(P) = P = \frac{10}{25}$

Event P has just occurred

P (Q) is a conditional expression, so

$P(Q|P) = \frac{9}{24}$

Now that we have used the multiplication rule, we can say that

$P(Q|P) = \frac{9}{24} \times \frac{9}{24} = \frac{3}{20}$

12.5 APPLICATIONS OF MULTIPLICATION THEOREM:

In probability theory and other mathematical contexts, the multiplication theorem has numerous significant applications in a variety of domains. The following applications are noteworthy:
1. Probability Theory: The multiplication theorem (or multiplication rule) is employed to determine the joint probability of two independent events. The multiplication theorem states that for two events, A and B,
$P(A \cap B) = P(A) \times P(B)$

Risk Assessment: Determine the probability of the simultaneous occurrence of multiple independent risks in the financial and insurance sectors.

Quality Control: Evaluate the likelihood of a product passing multiple independent quality evaluations.

Reliability Engineering: Evaluate the dependability of systems that contain numerous independent components.

2. Statistics: The multiplication rule for conditional probability is a

statistical tool that assists in determining the likelihood of an event in the presence of another event. The theorem stipulates that for events A and B:

$(A \cap B) = P(A) \times P(B \mid A)$

Bayesian Inference: Revise the probability of a hypothesis in light of new evidence.

Medical Diagnostics: Ascertain the probability of a disease based on the results of a test.

Market Analysis: Evaluate consumer behavior by determining the likelihood of purchasing a product based on prior purchases. 3. Machine Learning: The multiplication rule is essential for the development of probabilistic models, such as Bayesian networks and Hidden Markov Models, in the field of machine learning. Applications: Natural Language Processing: Determine the probabilities of word sequences in language models.

Spam Detection: Evaluate the probability of an email being spam by examining the presence of specific keywords.

User preferences are predicted by recommendation systems, which combine the probabilities of multiple factors.

4. Genetics: The multiplication formula is employed in the field of genetics to determine the likelihood of inheriting specific traits based on the independent assortment of DNA.

Applications: Genetic Disorder Prediction: Determine the likelihood that progeny will inherit genetic disorders from their parents.

Optimization of breeding strategies to attain desired characteristics in plants and animals is known as breeding programs.

5. Reliability Engineering: The multiplication rule is employed to evaluate the reliability of systems that contain multiple independent components.

System Design: Determine the overall reliability of intricate systems, including manufacturing processes or electrical infrastructure.

Maintenance Planning: Formulate maintenance schedules to guarantee system reliability by considering the likelihood of component failure.

6. Finance: The multiplication theorem is employed in the field of finance to assess the combined probabilities of a variety of independent market events.

Portfolio Management: Evaluate the probability of attaining specific investment objectives by considering independent market factors. Risk Management: Assess the likelihood of concurrent adverse market fluctuations.

7. Operations Research: The multiplication rule is a valuable tool in the analysis of the performance of systems and processes that involve multiple independent phases in operations research. Supply Chain Management: Determine the likelihood of timely delivery in the presence of multiple independent delays. Project Management: Determine the probability of project completion within a specified timeframe by analyzing the durations of individual tasks.

8. Information Theory: In information theory, the multiplication rule is employed to determine the mutual information and joint entropy of independent sources.

Data Compression: Evaluate the probabilities of data sequences to optimize encoding schemes.

Communication Systems: Develop error-correcting codes that are effective by comprehending the joint probabilities of transmitted signals.

The multiplication theorem is a versatile instrument that facilitates the computation of joint probabilities and is essential in a variety of practical applications in a variety of disciplines.

12.6 BINOMIAL THEOREM:

Binomial probability is the likelihood of achieving precisely x successes in n repeated trials in a binomial experiment, where there are two potential outcomes.

The binomial probability, denoted as $nCx \cdot px \cdot (1-p)n-x$, is calculated when the likelihood of success on an individual trial is p. The notation nCx represents the number of distinct combinations of x items chosen from a collection of n objects. Certain textbooks use the notation (nx) as an alternative to nCx.

It is important to understand that if p represents the chance of success in a single trial, then (1−p) represents the likelihood of failure in a single trial.

Example:

What is the probability of getting 6 heads, when you toss a coin 10 times?

In a coin-toss experiment, there are two outcomes: heads and tails. Assuming the coin is fair , the probability of getting a head is $\frac{1}{2}$ or 0.5 .

The number of repeated trials: $n = 10$

The number of success trials: $x = 6$

The probability of success on individual trial: $p = 0.5$

Use the formula for binomial probability.

$${}_{10}C_6 \cdot (0.5)^6 \cdot (1 - 0.5)^{10-6}$$

Simplify.

$$\approx 0.205$$

12.7 APPLICATIONS OF BINOMIAL THEOREM:

The binomial theorem, which offers a formula for expanding bino mial expressions raised to a positive integer power, has a wide rang e of applications in various disciplines.

The following are a few critical applications:

1. Applications of Algebra and Polynomials:

Polynomial Expansion: The binomial theorem enables the expansi on of expressions such as ($(a + b)^n$ into a sum involving terms of the form $a^k b^{n-k}$ with binomial coefficients.

Simplification: Facilitates the expansion of polynomial expressions , thereby facilitating the interaction with polynomials of a higher d egree.

2. Probability and Statistics:

Binomial Distribution: The binomial theorem is essential for the de velopment of the binomial distribution formula, which represents t he number of successes in a fixed number of independent Bernoulli trials.

Probability Calculations: Employed to determine the probabilities of a variety of outcomes in experiments that involve binomially dis tributed random variables.

3. Counting Problems: The binomial coefficients (nk) are employ ed in combinatorial problems to determine the number of methods for selecting k elements from a set of n elements.
Pascal's Triangle: The binomial coefficients are the components of Pascal's triangle, which has practical applications in algebra, proba bility, and counting.

4. Calculus:

Taylor and Maclaurin Series:
The series expansions for functions involving binomials are derive d using the binomial theorem, which is helpful in the approximatio n of functions and the solution of differential equations.

Integration and Differentiation:
Enables the computation of integrals and derivatives of binomial e xpressions.

5. Applications of Computer Science:
Algorithm Analysis: The binomial theorem is employed in the anal ysis of algorithms, particularly in the calculation of the time compl exity of recursive algorithms.

Cryptography: Involves mathematical structures that utilize the bin omial theorem to comprehend polynomial functions over finite fiel ds. 6. Engineering: Uses:

Signal processing is employed in the analysis and design of filters, as well as in the examination of waveforms.

Control Systems: Facilitates the examination of system stability an d response in control theory. 7. Physics: Applications:
Quantum Mechanics: The binomial theorem is employed in the co mputation of quantum states and operators.

Statistical Mechanics: Assists in the expansion and approximation of expressions that are associated with thermodynamic quantities a nd partition functions. 8. Applications of Economics and Finance:
Option Pricing Models: Employed in the binomial options pricing model, which offers a method for valuing options by taking into ac count various potential future stock prices.

Risk Assessment: Assists in the modeling and analysis of financial risks by employing binomial trees and other discrete probability m odels. 9. Applications of Biology and Genetics:

Genetic Probability: Utilized to determine the likelihood of genetic traits manifesting in progeny in accordance with Mendelian inherit ance patterns.

Population Genetics: Facilitates the prediction of genetic variation over generations and the modeling of allele frequencies.
10. Applications of Mathematical Proofs:

For the purpose of verifying statements that involve polynomial id entities and binomial coefficients, inductive proofs are frequently e mployed in mathematical induction.

Proofs of Inequality: Utilized to establish inequalities that involve binomial coefficients and sums.

A versatile and potent instrument with a wide range of applications , the binomial theorem is indispensable in a variety of disciplines, i ncluding finance, engineering, mathematics, and science.

12.8 LET US SUM UP

Probability theory is a robust mathematical framework that has widespread applicability in several commercial fields. The addition rule, multiplication rule, and binomial theorem are three essential ideas in probability theory that are vital for modeling and understanding uncertainties in business situations.

The addition rule of probability is used in the context of mutually exclusive occurrences, which are events that cannot occur concurrently. This principle is often used in business to guide decision-making procedures involving mutually exclusive alternative courses of action. For example, a corporation may be contemplating two distinct marketing tactics for the introduction of a product. Each approach's likelihood of success may be computed independently, and the addition rule can be used to estimate the overall probability of the product's success by combining the success probabilities of each method.

Conversely, the multiplication rule of probability is applicable for assessing the chance of numerous independent occurrences happening simultaneously. This is often seen in corporate situations that include sequential decision-making or procedures. In the context of supply chain management, the likelihood of effectively obtaining raw materials from various suppliers may be determined

by using the multiplication rule. This enables firms to evaluate the entire dependability of their supply chain.

Moreover, the multiplication rule plays a vital role in risk assessment. In the insurance sector, the multiplication rule is used to estimate the likelihood of numerous catastrophes, such as accidents or natural disasters, happening at the same time. This aids insurance firms in determining suitable rates and effectively managing their entire risk exposure.

The binomial theorem, a mathematical technique for increasing the powers of binomials, is used in probability theory to handle a certain number of independent trials, where each trial has an equal likelihood of success. This is especially pertinent in corporate settings when there are recurring attempts or tests with binary results, such as either achieving success or experiencing failure. The binomial distribution is often used in fields such as quality control, where it is used to estimate the likelihood of faulty items in a batch based on the chance of individual flaws.

These probability ideas are quite important in the process of making financial decisions. The addition rule is used to evaluate the probability of various financial outcomes, such as profit or loss. The multiplication rule is crucial in portfolio management, as it determines the likelihood of attaining a certain rate of return based on the performance of individual assets in the portfolio.

Furthermore, risk management in finance is strongly dependent on these concepts of probability. When examining the likelihood of a portfolio of loans defaulting, the multiplication formula is used to calculate the combined probability of default for each individual

loan. This helps financial organizations in making well-informed judgments about lending and allocating suitable reserves.

These probability concepts are used in marketing to evaluate the effectiveness of promotional efforts and product launches. The addition rule is used to calculate the cumulative chance of success when many marketing channels are utilized concurrently, but the multiplication rule is applied when assessing the odds of success for separate promotional operations.

To summarize, the use of supplementary, multiplication, and binomial theorem of probability in business is vast and varied. The probability ideas provide a structured and numerical framework for making decisions, evaluating risks, and planning strategies in several business fields, enhancing the quality and efficiency of company processes.

12.9 KEY WORDS:

Addition Theorem: States that given mutually exclusive events, the probability of the union of these events is equal to the sum of their individual probabilities.

Multiplication theorem: is used to determine the compound probability of two or more occurrences involving many tasks.

12.10 ANSWERS TO CHECK YOUR PROGRESS:

1. The probability of the occurrence of at least one of two events is determined using the ……..of probability.
Answer: addition theorem

2. The Addition Theorem of Probability states that given……., the probability of the union of these events is equal to the sum of their individual probabilities.

Answer: mutually exclusive events

3 ……..= P (A or B) =P (A) +P (B)

Answer: P (A∪B)

4The ……is employed in the analysis of algorithms, part icularly in the calculation of the time complexity of recursive algor ithms.

Answer: binomial theorem

5. ($A \cap$) = ……

Answer: ( ) × ( )

12.11 TERMINAL QUESTIONS:

Q.1.Explain Concept of Addition theorem of probability and its applications?

Q2. Explain Concept of Addition and multiplication theorem of probability and its applications?

Q3. Explain Concept of Binomial theorem and its applications probability?

UNIT 13: PROBABILITY DISTRIBUTION: CONCEPT AND APPLICATIONS OF BINOMIAL, POISSON AND NORMAL DISTRIBUTION.

Structure

13.0 Objectives

13.1 Introduction

13.2 Random Variables

13.3 Discrete Probability Distribution:

13.4 Importance of Probability Distribution

13.5 Types of Theoretical or Probability Distribution

13.6 Binomial Distribution

13.7 Poisson distribution

13.8 Normal Distribution:

13.9 Normal Distribution:

13.10 Significance of Normal distribution:

13.11 Properties of the normal distribution

13.0 OBJECTIVES

After studying this unit, you should be able to:

- Acquire knowledge of the Probability Distribution
- Comprehend the fundamental principles of a binomial distribution.
- Describe concept of Poisson distribution.
- Explain concept, significance and properties of the normal distribution

13.1 INTRODUCTION

The outcome of a random variable is ambiguous in a probability distribution. Realization is the term used to describe the observation of the outcome in this context. It is a function that converts the Sample Space into a Real number space, which is referred to as the State Space. They may be either discrete or continuous.

The probability Distribution of a Random Variable (X) shows how the Probabilities of the events are distributed over different values of the Random Variable. When all values of a Random Variable are aligned on a graph, the values of its probabilities generate a shape. The Probability distribution has several properties (for example: Expected value and Variance) that can be measured. It should be kept in mind that the Probability of a favorable outcome is always greater than zero and the sum of all the probabilities of all the events is equal to 1.

Probability Distribution is basically the set of all possible outcomes of any random experiment or event.

13.2 RANDOM VARIABLES

The sample space of the random experiment is the domain of the random variable, a real-valued function. It is denoted as X (sample space) = Real number.

It is imperative that we acquire an understanding of the concept of Random Variables, as we may not be exclusively concerned with the probability of an event, but rather with the quantity of events that are associated with the random experiment. The following example illustrates the significance of random variables:

What is the necessity of random variables?

Consider the coin tosses as an illustration. The initial step will involve the tossing of a coin to determine the probability. We will employ the letters H and T to represent "heads" and "tails," respectively.

In a distinct situation, let us assume that we are casting two dice and are interested in determining the probability of receiving two numbers that add up to 6.

Therefore, in both of these scenarios, it is necessary to first determine the number of occurrences of the desired event, known as Random Variable X, in the sample space. This information will be subsequently employed to calculate the Probability P(X) of the event. Therefore, Random Variables are of assistance. Initially, it is necessary to establish the mathematical definition of a random variable.

A probability distribution is a statistical function that delineates the all-possible values and likelihoods that a random variable can assume within a specified range. Probability distributions are classified into two primary categories:

13.3 DISCRETE PROBABILITY DISTRIBUTION:

This is applicable to discrete random variables, which are variables that can take on a countable number of distinct values. Examples consist of:

Binomial distribution:

Indicates the number of successes in a fixed number of independent Bernoulli trials with the same probability of success.

Poisson distribution:

Indicates the quantity of events that occur within a definite time or space interval, with the events occurring at a known constant mean rate and regardless of the time since the most recent event.

Continuous Probability Distribution:

This is applicable to continuous random variables, which are variables that have an infinite number of potential values within a specified range. Examples consist of:

Normal Distribution (Gaussian distribution):

A bell-shaped curve is formed by a continuous variable whose values are symmetrically distributed around the mean.

Exponential Distribution:

Defines the duration of time between events in a Poisson process, with a constant rate of occurrence.

13.4 IMPORTANCE OF PROBABILITY DISTRIBUTION

The foundation of contemporary statistics is the probability distribution. Its significance should be classified under the following categories:—

(1) Frequency distribution nature and trend estimation under specific assumptions and conditions, the nature and trend of the frequency distribution can be approximated using the probability distribution.

(2) The foundation of logical decisions theoretical distributions can be employed to analyze the phenomenon of risk and uncertainty, which is highly beneficial for the development of rational decisions.

(3) Forecasting-The probability distribution serves as the foundation for forecasting, projection, and prediction.

(4) Substitute for actual data—They may be employed as a substitute for the actual distribution when the latter is either prohibitively expensive or impossible to obtain.

(5) Sampling Test-Probability distributions are used as benchmarks to compare the actual frequency distributions and determine whether the difference is significant or solely a result of sampling fluctuations.

(6) The solution of a variety of daily life problems is facilitated by the probability distribution in a significant number of practical situations. A ready-made garments manufacturer, for instance, determines the volumes of different sizes based on the standard distribution. Whether or not the process is under control can be determined through quality control using the Poisson distribution. Similarly, a market researcher can employ the Chi-square distribution to determine the modifications in consumer behavior and reactions following the alteration in the product's nature.

To summarize, Merrill and Fox state that "The serve as benchmarks against which to compare observed distributions and act as substitutes for actual distributions when the latter are costly to obtain or cannot be obtained at all.". They offer decision-makers a logical foundation for making decisions and are beneficial for making predictions based on limited information or theoretical considerations. In summary, probability distributions are essential in the field of statistical theory.

13.5 TYPES OF THEORETICAL OR PROBABILITY DISTRIBUTION

The following three distributions are more popular:

i. Binomial Distribution

ii. Poisson distribution

iii. Normal Distribution

13.6 BINOMIAL DISTRIBUTION

James Bernoulli (1654-1705), a Swiss mathematician, is the name with which the binomial distribution is associated. Nevertheless, it was published in 1713, eight years after his passing. It is also referred to as Bernoulli's Distribution. 'Binomial' is a term that refers to two dichotomous alternatives or categories, including trial and process. Therefore, in this distribution, frequencies are divided based on two aspects or two potential outcomes, which are referred to as "success" and "failure" for the sake of expediency.

Definition of Binomial Distribution:

The binomial distribution is a discrete frequency distribution that is founded on dichotomous alternatives, specifically the probability of data of success (desired event) and failure. This distribution can be represented as a probability density function as follows:

$$P_{(x)} = {}^nC_x \, q^{n-x} \, p^x$$

or $$P_{(r)} = {}^nC_r \, q^{n-r} \, p^r$$

whereas, p = Probability of success,

q = Probability of failure or $1 - p$

n = number of trials, x or r = number of successes in n trial

Conditions for the application or assumptions of the Binomial Distribution: The Binomial Distribution can only be applied under the following conditions:-

(1) Finite number of trials-In this distribution, the number of trials should be fixed and finite. Specifically, an experiment is repeated under identical conditions for a fixed number of tries.

(2) Mutually exclusive outcomes: In each trial, there must be only two mutually exclusive possible outcomes of the event. For instance, when we flip a coin, there are only two potential outcomes: Head or Tail, and only one of them must occur.

(3) Consistent probability in each trial: The probability of the event (or success) occurring in each trial, denoted by p, remains constant. For instance, the likelihood of a head or tail event remains constant across all unbiased coin tosses.

(4) Independent trials -- All trials must be independent of one another, meaning that the outcome of any trial should not be influenced by the outcome of a previous or consecutive trial.

(5) Discrete variable: The variable must be discrete, meaning that the outcome should be expressed in whole numbers, whether it is success or failure.

Binomial Distribution Characteristics or Properties: The primary characteristics of the binomial distribution are as follows:

(1) Theoretical frequency distribution - The binomial distribution i s a theoretical frequency distribution that is derived from the Berno ulli theorem of algebra.

(2) Discrete Probability Distribution-

The binomial distribution is a discrete probability distribution in w hich the number of successes is expressed in whole numbers, not in fractions, and the values of n are set to 0, 1, 2, 3, and so forth.

(3) Line graph presentation-

The binomial distribution can also be represented graphically throu gh a line graph, in which the X-
axis represents the number of triumphs and the Y-
axis represents the probability of success.

(4) Binomial Distribution Shape-

The binomial distribution's shape is contingent upon the values of p and q.

For example, the binomial distribution will be precisely symmetric al if both p and q are equal.

The binomial distribution will be asymmetrical if p and q are not e qual (p + q). (5) Main Parameters-

The binomial distribution is characterized by two primary paramet ers, p and q, which can be used to calculate the entire distribution. (6) Mean, S.D., and Variance: The mean, standard deviation, and v ariance of a binomial distribution can be determined using the follo wing formulas:

Mean = np

S.D. = $\sigma = \sqrt{npq}$

Variance = $\sigma^2 = npq$

(7) Applications of the Binomial Distribution-

The binomial distribution is employed in events that have a dichoto mous classification, such as the flinging of coins or the hurling of d ice.

(8) Sequence of p and q:

The expression shall be as follows, with the number of successes ar ranged in either ascending or descending order depending on the se quence:- 0, 1, 2,..... n = $(q+p)^n$,..... n..... 2, 1, 0 = $(p + q)^n$

(9) Expected frequencies—

The expected frequencies can be determined by multiplying N by t he probabilities of the binomial distribution. The following express ion is employed for this purpose:

N $(q+p)^n$ or N $(p + q)^n$

(10) Constant values—

The following formulae are employed for a variety of constant valu es in a binomial distribution:-

The first moment, $\mu 1$, is equal to zero

The second moment, $\mu 2$, is equal to npq.

Third Moment $\mu 3 = npq (q-p)$

Fourth Moment $\mu 4 = 3n^2p^2q^2 + npq (1 – 6pq)$

The mean of X is equal to np.

S.D. = $\sigma = \sqrt{npq}$

13.7 POISSON DISTRIBUTION

Simen Denis Poisson, a French mathematician, developed the Pois son distribution in 1837.

It is a probability distribution that is discrete.

The Poisson distribution is employed in situations where the value of p is extremely small, i.e., p approaches zero ($p \rightarrow 0$), and the val ue of n is extremely large. In these cases, the binomial distribution fails to provide the appropriate theoretical frequencies, and the pois son distribution is found to be much more suitable.

It is important to note that the Poisson distribution is a limiting for m of the binomial distribution. As n and p approach infinity and ze ro, respectively, the mean (np) remains constant and finite.
The behavior of uncommon events, such as the number of bacteria in a single drop of purified water, the number of printing errors per page, and the number of telephone calls arriving per minute at a te lephone switchboard, is described by the Poisson distribution.
Applications of the Poisson distribution:

In a variety of disciplines, the Poisson distribution has been emplo yed to model infrequently occurring events in terms of time (know n as Temporal Distributions), area (known as Spatial Distributions) , volume, or similar units.

The Poisson distribution may be implemented in the following sce narios:- (1) In insurance problems, to count the number of casualtie s,

(2) In determining the number of deaths due to suicides or rare dise ases,

(3) The number of typographical errors per page in a typed materia l or the number of printing mistakes per page in a book,

(4) In biology, to count the number of bacteria,

(5) In statistical quality control, to count the number of defects per item, (6) In waiting-

time problems, to count the number of incoming telephone calls or incoming customers.

(7) The daily number of accidents that occur on a congested road,

(8) The act of counting the number of particles that are emitted fro m a radioactive substance in the field of physics.

In summary, the Poisson distribution elucidates the behavior of dis crete variables in which the probability of the event occurring is ex ceedingly low and the total number of potential cases (or trials) is a dequately large.

Characteristics and Properties of the Poisson Distribution—

The primary attributes of the Poisson distribution are as follows:-

(1) Distribution nature-

The Poisson distribution is a discrete probability distribution in wh ich the number of successes is specified in whole integers, such as 0, 1, 2, 3,..... (2) Applications of Poisson distribution-

The Poisson distribution is employed in situations where the value of p is extremely small (i.e., p → 0), the value of q is nearly equal t o 1 (q→ 1), and the value of n is extremely large.

In other words, this distribution is deemed suitable for describing t he behavior of uncommon events. (3) Main parameter- The mean (m = np) is the primary parameter of the Poisson distrib ution.

The entire distribution can be constructed if the mean value is kno wn. (4) Distribution shape—

The Poisson distribution is invariably positively biased. Nevertheless, the distribution shifts to the right and skewness is di minished as the value of m increases.

(5) Constant values—

The constant values in the poisson distribution are as follows:-

(i) Arithmetic Mean or $\overline{X} = m = np$

(ii) Standard Deviation $\sigma = \sqrt{m} = \sqrt{np}$

(iii) Variance or $\sigma^2 = m = np$

(iv) $\mu_1 = 0$ (v) $\mu_2 = m$

(vi) $\mu_3 = m$ (vii) $\mu^4 = m + 3m^2$

(viii) $\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{m^2}{m^3} = \frac{1}{m}$ (ix) $\beta_2 = \frac{\mu^4}{\mu_2^2} = \frac{m + 3m^2}{m^2} = 3 + \frac{1}{m}$

(x) $\gamma_1 = \sqrt{\beta_1} = \frac{1}{\sqrt{m}}$ (xi) $\gamma_2 = \beta_2 - 3 = 3 + \frac{1}{m} - 3 = \frac{1}{m}$

(6) Distribution assumption: The Poisson distribution is predicated on the following assumptions:

(i) The probability of the occurrence or non-occurrence of an event does not affect the other events. (ii) The pro bability of the occurrence of more than one event in a very small in terval is negligible. (iii) The probability of success for a small spa ce or short interval of time is proportional to the space or length of the time interval, as the case may be.

13.8 NORMAL DISTRIBUTION:

Abraham DeMoivre, an English mathematician, was the first to dis cover the normal distribution in 1733. However, the practical impl ementation of the distribution is credited to French mathematician Laplace and German astronomer Karl Gauss.

In recognition of Gauss, the normal distribution is occasionally ref erred to as the Gaussian distribution.

Quetelet, Galton, and Fisher subsequently enhanced and implemen ted this distribution. Definition of Normal Distribution: The normal distribution is a continuous probability distribution in which the relative frequencies of a continuous variable are distribut ed in accordance with the normal probability law.

To put it simply, it is a symmetrical distribution in which the frequ encies are distributed uniformly around the mean of the distributio n. Describing the concept of normal distribution.

Ya Lun

Chou has noted that the normal curve is "the perfectly smooth and symmetrical curve that results from the expansion of the binomial (p + q)" as n approaches infinite.

The normal curve can be regarded as the point at which the binomi al distribution approaches infinity as n increases.

Alternatively, we could assert that the normal curve denotes a conti nuous and infinite binomial distribution or a standard normal distri bution.

13.9 NORMAL DISTRIBUTION:

Normal Distribution Assumptions—

The normal distribution is predicated on the following assumptions :-

(1) Independent causes: The factors or causes that influence events are not interdependent.

(2) Multiple causation-

The causal forces are numerous, and all causes are of approximatel y equal importance.

(3) Symmetrical: The causal forces are such that the maximal frequ encies are concentrated near the arithmetric mean.

Additionally, the magnitude and quantity of the deviations above a nd below the population mean are equidistant.

In other words, the number and magnitude of the deviations from t he mean on either side are equivalent.

The normal distribution is the most significant and practical theoret ical frequency distribution.

It serves as the foundation of contemporary statistics.

Merrit and Fox have stated that "If a statistician were to choose onl y one distribution to work with during their lifetime, they would al most certainly choose the normal distribution."

Despite the possibility that modern statisticians and applied econo mists could function without a computer, it would be exceedingly c hallenging to do so without the normal distribution.

13.10 SIGNIFICANCE OF NORMAL DISTRIBUTION:

The significance of the distribution is evident in the following instances:—

(1) Universality—

This distribution is universal because the frequency distribution is normal in almost all domains, with the exception of specific conditions.

The scientist Sir Francis Galton wrote, "I know of scarcely anything so apt to impress the imagination as the wonderful form of cosmic order expressed by the law of frequency of Error."

W. J. Youden, a renowned statistician, has expressed his admiration for the normal distribution in an artistic manner (symmetrical), as illustrated below:-

THE NORMAL

LAW OF ERROR

STANDS OUT IN THE EXPERIENCE OF MANKIND

AS ONE OF THE BROADEST GENERALIZATION OF NATURAL PHILOSOPHY. IT SERVES AS THE

GUIDING INSTRUMENT IN RESEARCHES

IN THE PHYSICAL AND SOCIAL SCIENCES AND

IN MEDICINE AGRICULTURE AND ENGINEERING.

IT IS AN INDISPENSABLE TOOL FOR THE ANALYSIS AND THE

INTERPRETATION OF THE BASIC DATA OBTAINED BY OBSERVATION EXPERIMENT.

(2) The characteristics of a normal distribution, including the height of adults and the length of foliage on a tree.

Consequently, the normal distribution is extensively employed in the investigation of natural phenomena.

(3) Binomial and Poisson Distribution Approximation-
The normal distribution is a reliable approximation to the binomial and Poisson distributions, particularly as the number of observatio ns increases.

It is worth noting that the computation of probability for discrete di stributions becomes challenging for large values of n. In such cases , the normal distribution can be employed with great ease and conv enience.

(4) Sampling Distribution Conformity-

For large degrees of freedom, the distributions of almost all exact s amplings, such as the Student's t-distribution, Fisher's z-distribution, Snedecor's F-distribution, and Chi-square distribution, conform to the normal distribution.

(5) Basis of small samples—

The fundamental presumption of the small sample theory is that th e source population from which the samples are derived follows a normal distribution.

(6) Statistical quality control-

This distribution is highly beneficial in the establishment of control limits in statistical quality control.

(7) The central limit theorem states that the normal distribution has the property of determining the limits of population values.
This theorem allows us to determine the upper and lower bounds o f a population value.

For instance, 99.73% of the items are covered within a range of po pulation mean ± 3 S.D.

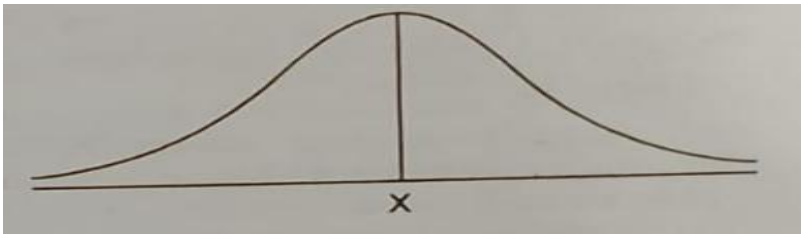
13.11PROPERTIES OF THE NORMAL DIST RIBUTION

The following are the significant properties of the normal distri bution or normal curve:- (1) Bell-

shaped: The normal curve is precisely symmetrical and bell-

shaped at the mean.

This implies that the two halves would coincide if the curve were t o be folded along its vertical axis at the center, as illustrated in the diagram.



(2) Continuous distribution-

The normal distribution is a distribution that encompasses continuo us variables.

Therefore, it is referred to as a continuous probability distribution.
(3) Central value equality - In a normal distribution, the mean, med ian, and mode are all equal.

(4) Uni-apex-The normal curve is uni-
apex because there is only one maximal point.
It is additionally referred to as a "Uni-
modal distribution" due to its singular mode.

(5) Equidistant Distance of Quartiles from Median-
In a normal distribution, the quartiles, which are denoted as Q1 and Q3, are equally distant from the median. Q3-M=M-Q1 as a result of this property.

(6) Asymptotic to the base line: The normal curve is asymptotic to the base line on both sides.

This implies that the base line is never touched by a normal curve, despite its propensity to do so.

The curve is expected to intersect the X-axis at infinity and becomes parallel to it at both extremities.

(7) Distribution Parameters: The mean (X) and standard deviation (S.D.) are the two primary parameters of a normal distribution.

These two parameters can be used to determine the entire distributi on. (8) Relationship between Q.D. and S.D.-

In a normal distribution, the quartile deviation (Q.D.) is 2/3 times t he standard deviation (S.D.), or 0-6745. In other words, Q.D. = S.D.

(9) **Relationship between M.D. and S.D.**—In a normal distribution, the mean deviation (M.D.) is 4/5 times the standard deviation, *i.e.*, M.D. $= \frac{4}{5}$ S.D.

(10) **Q.D., M.D. and P.E.**—In a normal distribution, Q.D. is equal to Probable Error, so P.E. = Q.D. Moreover, P.E. is 5/6 of M.D.

(11) **Constant values of Normal distribution**—In a normal distribution the values of different constants are as follows:—

(i) Mean $= \overline{X}$ or μ (ii) Standard Deviation $= \sigma$

(iii) First Moment $= \mu_1 = 0$ (iv) Second Moment $= \mu_2 = \sigma^2$

(v) Third Moment $= \mu_3 = 0$ (vi) Fourth Moment $= \mu_4 = 3\mu_2^2 = 3\sigma^4$

(vii) $\beta_1 = \frac{\mu_3^2}{\mu_2^3} = 0$ (viii) $\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{3\mu_2^2}{\mu_2^2} = \frac{3\sigma^4}{\sigma^4} = 3$

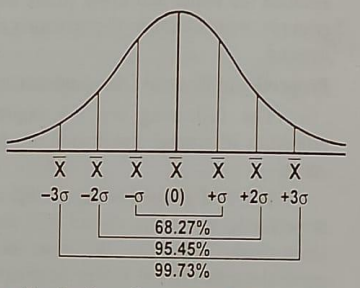
In a normal distribution, the standard value of β_2 is 3. If it is 3, the curve is called *mesokurtic*. If the value is more than 3, it will be *platykurtic* curve and in case of value less than 3 it will be called *leptokurtic*.

(12) **Area relationship**—The area under the normal curve, as shown in the following figure, is distributed as follows:—

(i) $\overline{X} \pm 1\ \sigma$ covers 68·27% frequencies (or area) 34·135% area will lie on either side of the mean,

(ii) $\overline{X} \pm 2\ \sigma$ covers 95·45% frequencies; 47·25% area lies on either side of the mean,

(iii) $\overline{X} \pm 3\ \sigma$ covers 99·73% frequencies; 49·865% are lies on either side of the mean.



The above figure exhibits the percentage distribution of area at a distance of 1 σ, 2 σ and 3 σ from the mean ordinate. The following table gives areas under various other distances:—

Distance from Mean ordinate	% of Total Area	Distance from Mean ordinate	% of Total Area
0·1 σ	3·983	1·0 σ	34·134
0·2 σ	7·926	1·5 σ	43·319
0·3 σ	11·791	1·96 σ	47·500
0·4 σ	15·542	2·00 σ	47·725
0·5 σ	19·146	2·5 σ	49·379
0·6 σ	22·804	2·58 σ	49·500
0·7 σ	25·804	3·0 σ	49·865
0·8 σ	28·814	3·5 σ	49·977
0·9 σ	31·594	4·0 σ	49·997

Thus, the two ordinates at distance 2·58 σ from the mean on either side would enclose 49·5 + 49·5 = 99% of total area.

13.12 LET US SUM UP:

A probability distribution is a mathematical function that depicts the likelihood of various potential values of a variable. Graphs or probability tables are frequently employed to illustrate probability distributions.

A probability distribution is a frequency distribution that is idealized.

A particular sample or dataset is described by a frequency distribution. It is the quantity of occurrences of each potential value of a variable in the dataset.

The probability of a value occurring in a sample is the determinant of the number of times it occurs. Probability is a numerical value between 0 and 1 that indicates the likelihood of an event occurring:

The value of zero indicates that the task is unattainable.

One indicates that it is certain.

In a sample, the frequency of a value increases as its probability increases.

To be more precise, the probability of a value is the relative frequency of that value in an infinitely large sample.

Probability distributions are theoretical in nature, as infinitely large samples are unattainable in actual life. They are idealized representations of frequency distributions that are designed to characterize the population from which the sample was selected.

The populations of real-world variables, such as the weight of poultry eggs or coin outcomes, are described using probability distributions. Additionally, they are employed in hypothesis testing to ascertain p-values.

13.13 KEY WORDS:

Probability distribution: is a mathematical function that depicts the likelihood of various potential values of a variable.

Normal curve: is precisely symmetrical and bell-shaped at the mean.

Poisson distribution: is employed in situations where the value of p is extremely small.

Binomial distribution: is a discrete frequency distribution

13.14 ANSWERS TO CHECK YOUR PROGRESS:

1. The total of the probabilities of all events in a probability distribution must equal …….

Answer: 1

2. A ……..distribution explains the likelihood of achieving one of two outcomes in a specific number of trials.

Answer: binomial

3. The ……distribution is used to represent the frequency of occurrences happening within a certain timeframe or area.

Answer: Poisson

4. In a ……distribution, about 68% of the data falls within one standard deviation of the mean.

Answer: normal

The form of a normal distribution is symmetrical.

Answer:

5. The ……..distribution is a continuous probability distribution with parameters μ (mean) and σ (standard deviation).

Answer: normal

13.15 TERMINAL QUESTIONS:

1. What is meant by theoretical frequency distribution? Describe the chief characteristics of Binomial, Normal and Poisson Distribution.
2. What do you understand by theoretical frequency distribution? Discuss in brief the various types of theoretical frequency distribution.

3. Distinguish between the normal and binomial distribution and discuss briefly the importance of normal distribution.

4. What is meant by theoretical distribution? Describe the chief characteristics of Binomial Distribution.

5. What are Binomial Distribution and Poisson distribution? Distinguish between them.

BLOCK-V: ESTIMATION THEORY AND HYPOTHESIS TESTING

UNIT 14: SAMPLING THEORY

Structure

14.0 Objectives
14.1 Introduction
14.2 Population:
14.3 Sample
14.4 Advantages of Sampling:
14.5 Limitations of sampling
14.6 Applications of Sampling:
14.7 Sampling Techniques:
14.8 Probability Sampling:
14.9 Types of Probability Sampling
14.10 Non-Probability Sampling
14.11 Types of Non-Probability Sampling
14.12 Opting for the Appropriate Methodology
14.13 Let Us Sum Up
14.14 Key Words
14.15 Answers to Check Your Progress
14.16 Terminal Questions

14.0 OBJECTIVES

After studying this unit, you should be able to:

• Understand the idea of sampling and its significance in statistical analysis

• Comprehend and use both probability and non-probability sampling methodologies.

• Acquire the knowledge of dividing the population into distinct groups and selecting samples from each category to guarantee the inclusion of various subcategories.

• Utilize sampling theory to create surveys and tests that provide dependable and accurate outcomes.

14.1 INTRODUCTION

Sampling theory is a fundamental aspect of statistics that pertains to the selection of a subset (a sample) from a larger population in order to estimate the characteristics of the entire population. The primary objective of sampling theory is to guarantee that the sample accurately represents the population, thereby enabling the drawing of valid inferences and conclusions. In sampling theory, the following are the primary concepts:

14.2 POPULATION:

In statistics and research, the terms "population" and "universe" are frequently used interchangeably to denote the comprehensive collection of items or individuals that are the subject of a specific study. A more comprehensive explanation of each is provided below:

In the field of statistics, a population is the comprehensive collection of individuals or items that are the subject of study and the basis for drawing conclusions.

Examples: When conducting a survey on local health practices, the entire population of a city is surveyed.

When conducting an analysis of defect rates, all vehicles manufactured by an organization during a particular year are considered.

In the context of research, it is common for researchers to gather data from a sample, which is a smaller, more manageable portion of the population. The results of the sample are subsequently extrapolated to the entire population.

Universe Definition: The term "universe" is frequently employed in the context of a more abstract or theoretical set of all possible observations or outcomes, and it is similar to "population."

Examples: The potential results of tossing a die.

The total number of potential responses to a survey query in a hypothetical scenario in which the entire globe is surveyed.

In the context of research, the universe may encompass all prospective elements that could be incorporated into a study, offering a more expansive scope than a population, which is frequently more precisely defined.

Key Points

Comprehensiveness: Both terms comprise all potential subjects or items within a predetermined scope.

Scope: The population is typically employed in practical, concrete contexts (e.g., a specific group of individuals), whereas the universe can be employed in more theoretical contexts (e.g., all possible outcomes).

It is essential to comprehend these terms in order to make precise inferences in research and statistics, to design studies, and to sample.

14.3 SAMPLE:

In the field of research and statistics, a sample is a subset of the population that is chosen for the purpose of the study. The primary objective of employing a sample is to draw conclusions about the population without the necessity of studying each individual, which is frequently impractical or impossible due to constraints such as time, cost, and accessibility. Here is a comprehensive examination of the concept of a sample:

A sample is a subset of a larger population that is selected to accurately represent the population. This subset is scrutinized and analyzed in order to derive conclusions that can be applied to the entire population.

Objective: Sampling enables researchers to more effectively acquire and analyze data. Researchers can conserve resources by inferring trends, patterns, and relationships within the population through the examination of a sample. Additionally, it assists in the formulation of predictions and decisions that are predicated on the data that has been collected.

14.4 ADVANTAGES OF SAMPLING:

Cost-effective: Conducting a comprehensive study of a complete population can be both time-consuming and costly. By concentrating on a manageable portion of the population, sampling reduces costs.

Time-Saving: The collection of data from a sample is more efficient than surveying the entire population, enabling the collection and analysis of data to be completed more quickly.

Practicality: In numerous instances, it is either impossible or impractical to conduct a comprehensive study of the entire population, particularly when the population is excessively large or dispersed. Sampling is a viable alternative.

Handling a lesser dataset simplifies the process, analysis, and management of the data, particularly when resources are restricted.

Precision and Accuracy: By employing appropriate sampling methods, samples can generate precise and accurate estimates of the parameters of the population. This is particularly true when the sample is representative of the population.

Reduced Data Overload: The use of a sample prevents data overload, which can occur when working with large datasets. This assists in concentrating on critical insights without becoming inundated by the sheer volume of data.

Longitudinal Studies' Feasibility: Samples are more suitable for longitudinal studies, which entail recurrent observations of the same variables over time. It is simpler to monitor a sample over time than to monitor an entire population.

Ethical Considerations: In certain research fields, such as medical studies, it is unethical or hazardous to subject the entire population to experimental treatments. Sampling enables researchers to conduct studies in a safe and ethical manner.

Flexibility: Sampling provides researchers with the ability to select from a variety of sampling methods (e.g., random, stratified, cluster) to align with their research objectives and constraints, thereby facilitating the design of their studies.

Improved Data Quality: By concentrating on a smaller group, the quality of the data collected is more detailed and of higher quality, thereby enhancing the overall quality of the research findings.

Sampling is a potent instrument for data analysis and research, as it facilitates the efficient, practicable, and precise examination of populations.

14.5 LIMITATIONS OF SAMPLING

Although sampling provides numerous benefits, it also has a number of drawbacks:

Sampling bias: Results may be biased if the sample is not representative of the population. This may be the result of non-random selection, under coverage, or over coverage of specific categories.

Sample Error: The estimation of population parameters from a sample is invariably subject to a certain degree of error. The veracity of the results can be influenced by this error, which is inherent in the sampling procedure.

Non-Response Bias: Non-response bias may result when individuals chosen for the sample fail to participate. The findings may be distorted if the non-respondents exhibit a substantial difference from the respondents.

Generalization Challenges: The results of a sample may not always be applicable to the entire population, particularly if the sample is limited or not properly selected.

Sampling Design Complexity: The development of a suitable sampling plan can be intricate and necessitates the careful consideration of a variety of factors, including the size of the sample, the sampling procedure, and any potential biases.

Limited Detail: The variability and subtleties that are present in the entire population may not be captured by the samples, resulting in a loss of detailed information.

A sample's accuracy is contingent upon the quality of the sampling frame, which is the list from which the sample is selected. An unrepresentative sample may be the consequence of an outmoded or defective sampling frame.

Cost and Resource Constraints: Although sampling is generally cost-effective, certain sampling methods (e.g., stratified or cluster sampling) can be resource-intensive and necessitate additional expert knowledge and effort.

Ethical and Legal Challenges: In certain instances, the acquisition of a sample may be accompanied by ethical or legal challenges, including concerns regarding data protection, consent, and privacy.

Limitations of Sampling Techniques: Each sampling technique has its own set of constraints. For example, random sampling may not always be feasible, and stratified sampling necessitates precise stratification identification, which may not always be feasible.

Sample Data Misinterpretation: The analysis and reporting of results may result in the misinterpretation of sample data if the limitations and potential biases of the sampling procedure are not properly acknowledged and addressed.

Sample Size Determination: The determination of the appropriate sample size is essential and can be difficult. An insufficiently sized sample can result in unreliable results, while an excessively large sample can be unnecessarily time-consuming and expensive.

In order to design studies that are robust, interpret findings accurately, and make informed decisions based on sample data, it is essential for researchers to comprehend these limitations.

14.6 APPLICATIONS OF SAMPLING:

Sampling is a fundamental technique that is employed in a variety of disciplines to draw conclusions about a population by examining a smaller, more manageable subset of that population. Selected implementations of sampling across various disciplines are as follows:

1. Statistics and Research Survey Research: Sampling enables researchers to collect data from a subset of a population in order to derive conclusions about the entire population. This is a frequently employed method in market research and public opinion surveys.

Experimental Studies: Clinical trials involve the selection of a sample of patients to evaluate the efficacy of a novel treatment or substance.

Quality Control: Sampling is employed by manufacturers to evaluate the quality of products in a production line without the necessity of inspecting each individual item.

2. Machine Learning and Data Science Training Models: In order to make predictions about a wider dataset, machine learning models are frequently trained on a sample of data.

Big Data Analysis: Sampling can be employed to enhance the feasibility and efficiency of data analysis when dealing with large datasets.

3. Environmental Science

Ecological Studies: In order to evaluate biodiversity, monitor environmental changes, and investigate the distribution and abundance of species, scientists sample specific areas.

Pollution Monitoring: The process of detecting and quantifying pollution levels by sampling air, water, and soil.

4. Social Sciences

Sociological Studies: In order to comprehend social phenomena, sampling methods are employed to examine a representative group within a larger population.

Anthropology: Fieldwork frequently entails the surveying of particular groups or communities in order to gain insight into social structures and cultural practices.

5. Medicine and Healthcare Epidemiology: Sampling is essential for the analysis of the spread of diseases, the investigation of risk factors, and the assessment of the efficacy of interventions.

Health Surveys: Sampling is employed in national health surveys to collect data on health behaviors, conditions, and access to healthcare services.

6. Economics Market Analysis: Economists employ sampling to evaluate economic indicators, consumer behavior, and market trends.

Census and Demographics: In order to offer comprehensive demographic insights, sampling methods are frequently implemented in conjunction with census data.

7. Agriculture Crop Surveys: Sampling techniques are employed to monitor parasite infestations, assess soil quality, and estimate crop yields.

Animal Studies: The analysis of reproductive patterns, health status, and population dynamics is facilitated by the sampling of animal populations.

8. Engineering Reliability Testing: Engineers assess materials and components to evaluate their durability and dependability.

Process Optimization: Sampling is a method that assists in the monitoring and enhancement of industrial processes by analyzing a subset of outputs.

9. Marketing Product Testing: Prior to a full-scale launch, companies implement sampling to evaluate new products with their intended audience.

Consumer Feedback: The process of gathering customer feedback in order to enhance the quality of products and services.

10. Finance Auditing: Financial auditors employ sampling to assess the veracity of financial statements by examining a subset of transactions and accounts.

Risk Management: Financial risk assessments employ sampling to simulate potential losses and inform investment decisions.

Sampling is intended to save time, reduce costs, and make the analysis more practicable by allowing for the collection of insights and the formulation of decisions based on a smaller, more manageable portion of the population in each of these apps.

14.7 SAMPLING TECHNIQUES:

Sampling techniques are employed to select a subset of individuals or objects from a larger population. The techniques can be broadly classified into two categories: probability sampling and non-probability sampling. The following is a comprehensive description of a variety of sampling methods:

14.8 PROBABILITY SAMPLING:

In probability sampling, each member of the population has a known, non-zero probability of being selected. This enables the application of the findings to the entire population.

14.9 TYPES OF PROBABILITY SAMPLING

1.Simple Random Sampling:

Simple random sampling is a fundamental sampling method in which each member of a population has an equal likelihood of being selected. It is a fundamental concept in statistics and is frequently employed to guarantee that a sample accurately represents the population. The process is as follows:

Define the Population: Identify the group of individuals or items from which you wish to obtain a sample. This population should be well-defined and finite.

Numbering: Assign a distinct number to each member of the population.

Random Selection: Employ a random method to select a sample from the population. This can be accomplished by employing the following:

Random Number Tables: Tables that contain random numbers.

Computers or calculators: Employ software or instruments that can generate random numbers.

extracting from a Hat: The act of physically extracting numbers from a container.

Sample Size: Determine the number of members you wish to include in the sample, ensuring that it is suitable for the analysis you intend to conduct.

For instance,

Suppose that you have a population of 1000 individuals and you wish to select a sample of 100 individuals:

Define the Population: The population consists of 1,000 individuals.

Assign Numbers: Assign a number between 1 and 1000 to each individual.

Random Selection: Employ a random number generator to generate 100 distinct numbers between 1 and 1000.

Sample Selection: The sample is composed of the individuals who correspond to the 100 numbers.

Benefits

Simplicity: Simple to comprehend and execute.

Unbiased: Selection bias is mitigated by the fact that each member of the population has an equal chance of being selected.

Representative: The sample is likely to be representative if the population is homogeneous.

Negative aspects

Not Always Practical: May be challenging to implement with large or dispersed populations.

Complete inventory: It is essential to have a comprehensive inventory of the population, which may not always be accessible.

Homogeneity Assumption: Assumes that the population is homogeneous, which may not be the case in the actual world.

Utilizations

Surveys and polls: To guarantee that the sample is representative of the general population.

Quality Control: The process of arbitrarily testing products for defects in the manufacturing process.

Research Studies: In the medical, social, and other disciplines of research, the objective is to generalize the results to the general population.

2.Systematic Sampling

Systematic sampling is a probability sampling technique that involves the selection of elements from a larger population based on a predetermined, periodic interval. The sampling interval is determined by dividing the population size by the intended sample size. The following is a detailed procedure for the implementation of systematic sampling:

Define the Population: Clearly specify the population from which you intend to obtain a sample.

Sample Size Determination: Determine the number of elements you wish to incorporate into your sample.

Determine the Sampling Interval (k):

k **= Population Size/ Size of the Sample**

Round this number to the nearest whole number.

Randomly Choose a Starting Point: Select an arbitrary starting point between 1 and k. This guarantees that each element in the population has an equal opportunity to be selected at the outset.

To select sample elements, begin at the randomly selected beginning point and select every k-th element in the population until the desired sample size is achieved.

For instance,

Suppose that you have a population of 1,000 individuals and you wish to select a sample of 100 individuals.

Population Size (N): 1,000

Number of Samples (n): 100

Determine the value of k:

$k = 1000 / 100 = 10$

Random Starting Point: Assume that you arbitrarily select the number 8 as your starting point. 8 Sample Elements Selection: The 8th individual would be selected, followed by the 18th, 28th, 38th, and so on, until a total of 100 individuals are included in the sample.

Benefits of Systematic Sampling

Simplicity: It is more straightforward to execute than plain random sampling, particularly for large populations.

Efficiency: Requires a reduced amount of time and resources.

Coverage: Guarantees uniform coverage throughout the entire population.

Systematic Sampling's Drawbacks

Periodicity: Results may be biased if there is an underlying pattern in the population that corresponds with the sampling interval.

Lack of Randomness: Although the starting point is determined at random, the selection process adheres to a predetermined pattern, which may induce bias.

Systematic sampling is an extensively used method in a variety of disciplines, including market research, quality control, and environmental studies, due to its simplicity and effectiveness.

3. Stratified Sampling

Stratified sampling is a statistical research technique that is employed to guarantee that distinct subgroups within a population are adequately represented. Details regarding its operation are as follows:

Classify Strata: Divide the population into distinct subgroups or strata based on a specific characteristic, such as age, income, or education level.

Random sampling is implemented within each stratum. This can be accomplished through systematic sampling, simple random sampling, or any other sampling method.

Sample Combination: The aggregate sample is formed by combining the samples from each stratum.

Benefits:

Enhanced Precision: Stratified sampling can result in more precise and reliable estimates than simple random sampling by guaranteeing that each subgroup is represented.

Representation: It guarantees that significant subgroups are not disregarded.

List of disadvantages:

Complexity: The administration and analysis of this system can be more intricate, particularly when the population is large or the strata are numerous.

Information Needed: Accurate stratification necessitates comprehensive population knowledge.

This approach is especially advantageous when the population is heterogeneous and it is necessary to ensure that the sample accurately represents the diverse subgroups.

4. Cluster Sampling

Cluster sampling is another statistical sampling technique that is employed when it is impractical or challenging to compile a comprehensive inventory of the entire population. The process is as follows:

Divide the population into clusters: The population is typically divided into clusters based on natural or existent divisions, such as geographical areas, institutions, or households.

Randomly Select Clusters: Select a random sample of clusters. This can be accomplished through the use of simple random sampling or another sampling method.

Sample within Clusters: The method of sampling within clusters is contingent upon the study design.

Sample All Units within Selected Clusters: Incorporate each individual or unit within the selected clusters.

Sample within Clusters: Conduct additional sampling within the selected clusters.

Benefits:

Cost-Effective: Frequently more practicable and economical, particularly when the population is geographically dispersed.

Reduction in Travel: Clusters that are geographically located can reduce travel and logistical expenses.

List of disadvantages:

Reduced Precision: Clusters may be internally heterogeneous, which may result in a lower level of precision than other methods such as stratified sampling.

Potential Bias: The results may be biased if the clusters are not representative of the population.

Cluster sampling is advantageous in surveys that are conducted on a large scale or in which populations are organically organized, as it is impractical to develop a comprehensive inventory of the population.

5. Sampling with Multiple Stages

Multiple stage sampling is a sophisticated form of cluster sampling that entails the aggregation of multiple levels. If an exhaustive inventory of the entire population is unavailable or if a population is dispersed over a large area, this technique is frequently employed. The process is as follows:

First Stage - Primary Clusters: The population is divided into primary clusters, such as regions or localities. Select a random sample of these primary clusters.

Second Stage - Secondary Clusters: Within each primary cluster that has been chosen, identify secondary clusters, such as schools or localities. Select a random sample of these secondary clusters.

Subsequent Stages: If required, this process may be extended with additional stages. For example, households or individuals may be further sampled within the secondary clusters.

Data Acquisition: Collect data from the ultimate level of sampling units.

Benefits:

Efficiency: By emphasizing smaller, more manageable units, the cost and complication of data collection are diminished.

Practicality: Beneficial in situations where it is impractical to directly enumerate or sample the entire population due to its large size and geographical dispersion.

List of disadvantages:

Complexity: Enhances the intricacy of the data analysis and sampling process.

Potential for Enhanced Sampling Error: The potential for sampling error is increased by the introduction of variability with each additional stage of sampling.

In large-scale surveys, such as national health surveys or educational assessments, the population is frequently subdivided into a variety of subgroups, which is called multiple stage sampling.

14.10 NON-PROBABILITY SAMPLING

Non-probability sampling is advantageous for exploratory research, as it does not guarantee that each member of the population will be selected without bias.

14.11 TYPES OF NON-PROBABILITY SAMPLING

1. Sampling for Convenience

Convenience sampling is a non-probability sampling procedure that involves participants being chosen based on their proximity to the researcher and their ease of availability. The accessibility and cost-effectiveness of this approach make it a popular choice; however, it may introduce bias because the sample may not be representative of the broader population.

For instance, the sample may not accurately represent the experiences of students at other institutions or those who do not attend college if a researcher only surveys those in their own courses in order to study college students. Convenience sampling is

frequently implemented in exploratory research or when resources are restricted, regardless of these constraints..

2. Purposive or Judgmental Sampling

Purposive or judgmental sampling is a non-probability sampling procedure in which participants are chosen based on the researcher's judgment or specific criteria. This method is frequently employed by researchers who wish to concentrate on a specific subset of the population that satisfies specific characteristics or conditions.

For instance, purposive sampling may be implemented by a researcher who is investigating a rare medical condition to identify exclusively those who possess the condition. This approach is particularly advantageous for qualitative research that seeks to acquire comprehensive insights from a specific demographic.

Although purposive sampling enables a focused approach, it may introduce bias because the selection is dependent on the researcher's judgment rather than arbitrary sampling. This may restrict the generalizability of the results to a broader population.

3. Quota Sampling

Quota sampling is a non-probability sampling method that guarantees that specific subgroups of the population are present in the sample in proportion to their prevalence in the population. It entails the establishment of quotas for a variety of characteristics, such as age, gender, and ethnicity, and the subsequent selection of participants until those quotas are successfully met.

Identify Quotas: Determine the main characteristics and the percentage of the population that each subgroup represents.

Participants who are chosen: Recruit participants to fulfill the quotas for each subgroup. This can be accomplished through the use of convenience sampling or other methods.

For instance, a researcher may recruit participants until the desired demographic ratios are achieved, such as 50% male and 50% female, in order to conduct a survey of a community.

Although quota sampling can guarantee that various subgroups are represented, it does not incorporate randomization, which may introduce bias and reduce the generalizability of the results.

4. Snowball Sampling

Quota sampling is a non-probability sampling method that guarantees that specific subgroups of the population are present in the sample in proportion to their prevalence in the population. It entails the establishment of quotas for a variety of characteristics, such as age, gender, and ethnicity, and the subsequent selection of participants until those quotas are successfully met.

Identify Quotas: Determine the main characteristics and the percentage of the population that each subgroup represents. Participants who are chosen: Recruit participants to fulfill the quotas for each subgroup. This can be accomplished through the use of convenience sampling or other methods.

For instance, a researcher may recruit participants until the desired demographic ratios are achieved, such as 50% male and 50% female, in order to conduct a survey of a community.

Although quota sampling can guarantee that various subgroups are represented, it does not incorporate randomization, which may introduce bias and reduce the generalizability of the results. Snowball sampling is a non-probability sampling technique that is frequently employed to investigate populations that are difficult to access or concealed. This method entails an initial participant who subsequently refers the researcher to other prospective participants. As additional individuals are identified and enrolled in the study,

they, in turn, refer others, thereby generating a "snowball" effect. Begin with the initial participants: Initiate the study with a limited number of participants who satisfy the prescribed criteria. Referrals: Request that these initial participants refer others who also meet the criteria.

Sample Expansion: Continue the referral process until the desired sample size or saturation is achieved.

Research on niche groups or populations that are difficult to access, such as individuals with rare maladies or members of specific subcultures, is particularly beneficial when conducted using snowball sampling. Nevertheless, the sample is not randomly selected, and participants are likely to refer individuals who are similar to themselves, which may limit the diversity within the sample. Consequently, bias may be introduced.

5. Self-Selection Selection:

Self-selection sampling is a non-probability sampling method in which individuals opt-in or volunteer to participate in a study, rather than being randomly selected. This approach is frequently implemented in surveys, online research, and studies that facilitate participants' responses according to their availability or interests.

Invitation to Participate: Researchers frequently utilize advertisements, announcements, or requests for volunteers to solicit participation from individuals.

Volunteering is a voluntary activity in which individuals who are interested or meet specific criteria elect to participate.

For instance, a researcher may publish a survey link on social media, which is accessible to anyone who wishes to participate.

Although self-selection sampling can be cost-effective and practicable, it can introduce substantial bias. The sample may not be representative of the broader population, particularly if the motivations for volunteering are related to the study topic, as it is composed of individuals who voluntarily participated. This has the potential to influence the generalizability of the findings.

14.12 OPTING FOR THE APPROPRIATE METHODOLOGY

Research Objectives: Determine whether the objective is to investigate specific cases or to generalize the results to a population.

Population Characteristics: Evaluate the population's accessibility, diversity, and size.

Resources: Evaluate the available manpower, budget, and time.

Needed Precision: Determine the necessary level of confidence and accuracy in the results.

The selection of a sampling technique is contingent upon the specific objectives and constraints of the research, as each technique has its own advantages and disadvantages.

14.13 LET US SUM UP

Probability sampling and non-probability sampling are the two primary categories into which sampling methodologies can be broadly classified.

Probability Sampling: Each member of the population has a known, non-zero probability of being selected. This procedure guarantees that the sample is representative of the population, thereby enabling

the production of generalizable results. Types that are frequently encountered include:

Simple Random Sampling: Each member has an equal likelihood of being chosen. For instance, selecting names from a hat.

Stratified Sampling: The population is divided into subgroups (strata) based on specific characteristics, and samples are selected from each stratum. Assuring representation across critical subgroups is ensured.

Cluster Sampling: The population is divided into clusters, typically based on geography, and entire clusters are randomly selected. This technique is advantageous when the population is immense and dispersed.

Systematic sampling involves the selection of every nth member of the population. For instance, selecting the tenth individual from a list.

Non-Probability Sampling: The likelihood of each member being selected is unknown. This approach is frequently more convenient and expeditious; however, it may induce bias. Types that are frequently encountered include:

Convenience Sampling: Samples are selected based on their accessibility. For instance, conducting interviews with individuals at a purchasing facility.

Purposive or Judgmental Sampling: Samples are chosen according to the researcher's assessment of which individuals would be the most representative or beneficial.

Quota Sampling: The researcher guarantees that the sample adheres to specific quotas, such as the number of distinct demographic groupings.

The snowball sampling method involves the recruitment of prospective subjects from among their acquaintances by existing study subjects. This is frequently employed in research that involves populations that are difficult to access.

The Significance of Sample Size

The accuracy and reliability of the research findings are significantly influenced by the size of the sample. In general, a larger sample size results in more reliable and accurate results, which reduces the margin of error and increases the efficacy of the study. Nevertheless, the analysis of extremely massive samples can be resource-intensive. Consequently, the process of determining an optimal sample size entails the careful consideration of resource constraints, feasibility, and accuracy.

Bias and Sampling Error

Sampling Error: This phenomenon is the result of the inherent variations that arise when a sample is examined, as opposed to the entire population. It is possible to reduce its severity, but it cannot be entirely eradicated.

Sampling bias is the phenomenon in which designated members of the population are systematically more likely to be included in the sample than others, resulting in unrepresentative results. Through meticulous sampling design and execution, bias can be mitigated.

Utilization of Sampling in Research

Sampling is employed in a variety of disciplines, such as environmental studies, market research, healthcare, and social sciences. For example,

Sampling is a valuable tool in the healthcare industry, as it enables the testing of novel remedies on a smaller group before they are applied to the general population.

In market research, companies employ samples to better understand consumer preferences and behaviors, which in turn informs their marketing and product development strategies.

In the field of social sciences, researchers analyze samples of populations to gain insight into social behaviors, trends, and issues. This information is used to inform academic theories and policy decisions.

In conclusion, a sample is an essential instrument in research, enabling the efficient collection and analysis of data. The validity and reliability of the research findings are significantly influenced by the selection of the sampling method and sample size. The results are more valuable for decision-making and further research when they can be accurately generalized to the broader population, which is facilitated by the use of appropriate sampling techniques.

14.14 KEY WORDS:

Random sampling: is the use of randomization to pick a smaller subset of the population for examination and research.

Non-probability sampling: is one in which the probability of picking each person is not known.

14.15 ANSWERS TO CHECK YOUR PROGRESS:

1. A ..... is a portion of a population that is used to represent the full group

Answer: sample

2. The technique of picking a sample in which each individual in the population has an equitable probability of being picked is referred to as _____ sampling.

Answer: random

..... Sampling is the process of separating the population into subgroups and then selecting a sample from each category.

Answer: Stratified

4. In ……..sampling, researchers choose a sample based on their understanding of the population and the objectives of the study.

Answer: purposive

5. ..... is the discrepancy between the outcome of a sample and the actual outcome of the whole population, resulting from chance.

Answer: Random error

6. The process of inferring population characteristics from a sample is known as .....

Answer: statistical inference

7.The ..... is the whole set of persons or situations from whom we want to gather information.

Answer: population

14.16 TERMINAL QUESTIONS:

1. What is sampling and what are its objects? Discuss the various methods of selecting samples and indicate the cases when each one of them should be used.
2. "A good sample must be based on random selection".
3. Explain Types of Probability sampling techniques.
4. Explain Types of Non - Probability sampling techniques.

UNIT 15: ESTIMATION THEORY AND HYPOTHESIS TESTING

Structure

15.0 Objectives

15.1 Introduction

15.2 Varieties of Estimators

15.3 Common Methods of Estimation

15.4 Point Estimate:

15.5 Interval estimation:

15.6 Assumptions and Limitations of Interval Estimation

15.7 Hypothesis Testing

15.8 Hypothesis testing procedures

15.9 Commonly conducted tests

15.10 Let Us Sum Up

15.11 Key Words

15.12 Answers to Check Your Progress

15.13 Terminal Questions

15.0 OBJECTIVES

After studying this unit, you should be able to:

- Acquire knowledge of several categories of estimators, such as point estimators and interval estimators.
- Comprehend the process of constructing null and alternative hypotheses.
- Acquire knowledge about various test statistics and their use in decision-making.
- Comprehend the hypothesis testing procedures

15.1 INTRODUCTION:

Estimation theory is a subfield of signal processing and statistics that pertains to the estimation of parameter values using empirical or measured data. The fundamental objective of estimation theory is to derive the values of one or more parameters from a collection of observed data. It is essential in a variety of disciplines, including finance, economics, engineering, and the natural and social sciences.

Key Concepts in Estimation Theory Parameter: The unknown quantity that requires estimation.

Estimator: A rule or algorithm that generates an estimate of the parameter by analyzing the observed data.

Estimate: The precise numerical value that is determined through the use of an estimator.

Bias: The discrepancy between the genuine value of the parameter and the expected value of the estimator.

Variance is a metric that quantifies the extent to which the estimator's values deviate from its anticipated value.

Mean Squared Error (MSE): An estimate's expected value of the squared difference between the estimator and the true parameter value, which is a combination of the estimator's bias and variance.

Consistency: An estimator is consistent if it converges in probability to the true value of the parameter as the sample size increases.

Efficiency: An estimator that is efficient is the one with the least variance among all unbiased estimators.

Cramér-Rao Bound: Offers a lower bound on the variance of unbiased estimators of a parameter.

15.2 VARIETIES OF ESTIMATORS

Point Estimator: Offers a singular value estimate of a parameter.

Interval estimator: Indicates a confidence interval (range of values) within which the parameter is anticipated to fall, with a specific probability.

15.3 COMMON METHODS OF ESTIMATION

Maximum Likelihood Estimation (MLE) assesses the probability of the observed data given the parameters by maximizing the likelihood function, which is used to estimate the parameters.

The method of moments is used to estimate the parameters by equating the sample moments (e.g., sample mean, sample variance) to the theoretical moments of the distribution.

Least Squares Estimation: Reduces the sum of the squared differences between the predicted values and the observed values.

Bayesian Estimation: A posterior distribution is generated by combining prior information about the parameter with the observed data, from which estimates can be generated.

Utilizations

Signal Processing: The estimation of signals in the presence of noise.

Econometrics: The process of estimating economic models and parameters.

Control Systems: The estimation of system states and parameters.

Communication Systems: Prediction of transmitted signals and channel parameters.

Parameter estimation for algorithms and models in machine learning.

It is essential to comprehend estimation theory in order to make inferences and predictions based on data, thereby guaranteeing that the estimates obtained are as precise and reliable as possible.

15.4 POINT ESTIMATE:

Point estimation is a statistical technique that is employed to estimate an unknown population parameter by assigning a unique value (a "point") to it. The point estimate, which is a singular value, is derived from sample data and represents the most accurate estimate of the parameter. The following are the primary concepts associated with point estimation:

1. Estimator: The rule or formula that is employed to determine the point estimate. For instance, the sample mean (x¯\bar{x}x¯) is an estimator of the population mean (μ\muμ).

2. Point Estimate: The value that is derived by applying the estimator to the sample data. For instance, the point estimate of the population mean is the average of a sample of data points. 3. Qualities of Effective Estimators:

Unbiasedness: An estimator is unbiased if its expected value is equivalent to the parameter's true value. In other terms, it does not consistently overestimate or underestimate the parameter.

Consistency: An estimator is consistent if the point estimate converges to the true parameter value as the sample size increases.

Efficiency: An efficient estimator has the lowest variance among unbiased estimators. This implies that it offers the most accurate parameter estimate.

Sufficiency: An estimator is considered adequate if it incorporates all relevant data regarding the parameter.

Common Point Estimators:

1. Mean: The sample mean ($\bar{x}$) is frequently employed as a point estimate of the population mean (μ).

2. Proportion: The population proportion ((p^) is estimated using the sample proportion (p).

3. Variance: The sample variance (S^2) is employed as the point estimate of the population variance (σ^2).

For instance, Assume that you wish to determine the average height of adult males in a city. You randomly select a sample of 100 adult males and measure their heights. The average height of the sample is 175 cm. The population mean height is estimated to be 175 cm at this point. Utilization Point estimates are frequently employed in a variety of disciplines, including economics, biology, engineering, and social sciences, to derive conclusions about population parameters from sample data. They offer a straightforward and uncomplicated method of summarizing information; however, they fail to communicate the uncertainty associated with the estimate, which is where interval estimation and hypothesis testing are employed.

15.5 INTERVAL ESTIMATION:

Interval estimation entails the development of a confidence interval, which is a set of values that are likely to contain the unknown population parameter with a specific degree of confidence. For

instance, the true population mean would be present in 95% of all feasible samples within a 95% confidence interval for the population mean.

Interval Estimation Types

Depending on the parameter being estimated and the procedure employed to calculate the confidence interval, there are various varieties of interval estimation. Common varieties include:

Confidence interval for the population mean: This is employed to approximate the population mean in the absence of information regarding the population standard deviation. It calculates the critical value for the confidence interval by employing the t-distribution.The population proportion's confidence interval is as follows: This function is employed to determine the proportion of the population that supports a specific candidate, for example. It calculates the critical value for the confidence interval by employing the normal distribution.

Confidence interval for the difference between two means: This is employed to calculate the discrepancy between the means of two populations. Depending on the sample sizes and the assumption of equal variances, the t-distribution or normal distribution is employed.

Confidence interval for the difference between two proportions: This is employed to calculate the discrepancy between two population proportions. Depending on the sample sizes and the assumption of equal variances, the normal distribution or the chi-square distribution is employed.

15.6 ASSUMPTIONS AND LIMITATIONS OF INTERVAL ESTIMATION

Interval estimation, like point estimation, is subject to certain assumptions and has certain constraints. Some of the primary assumptions and constraints are as follows:

Interval estimation is predicated on the assumption that the sample is randomly selected and accurately represents the population of interest. If the sample is biased or non-random, the confidence interval may not be fully accurate.

Normality assumption: Interval estimation is predicated on the assumption that the population is normally distributed or that the sample size is sufficiently large to allow the central limit theorem to be applied. The confidence interval may be inaccurate if the data is not normally distributed and the sample size is small.

Independence assumption: Interval estimation is predicated on the assumption that the observations are independent of one another. If there is a correlation or dependence between the observations, the confidence interval may not be perfectly accurate.

Finite population correction: In the event that the sample size is a substantial proportion of the population size, it may be necessary to apply a finite population correction factor to modify the confidence interval.

In summary, point estimation and interval estimation are two critical statistical concepts that are employed to estimate population parameters using sample data. Point estimation yields a single value estimate for the population parameter, whereas interval estimation yields a range of values that the population parameter is likely to fall within. The selection of a method is contingent upon the specific research query and data available, as both have their own advantages

and limitations. The results must be interpreted appropriately, and it is crucial to comprehend the assumptions and conditions that are necessary for both methods.

15.7 HYPOTHESIS TESTING

Hypothesis testing is the process of using data to assess a hypothesis about a population. A hypothesis is a proposition that makes a claim about a characteristic of a population. For example, the null hypothesis states that the average value of the population is 5. A test statistic is a quantitative measure used to assess the validity of a hypothesis.

Hypothesis testing is a technique used to evaluate a claim or assumption about a population parameter.

15.8 HYPOTHESIS TESTING PROCEDURES

Hypothesis testing is a statistical method used to draw conclusions or make assumptions about a population based on the analysis of sample data. The procedure comprises a multitude of sequential stages:

1. Develop the hypotheses:

The null hypothesis (H0) states that there is no impact or difference. It represents a statement affirming the absence of change or the maintenance of the current state.

The alternative hypothesis (H_1 or H_a) states that there exists an effect or a difference. The purpose of your test is to achieve a certain goal.

2. Establish the level of significance (α):

The significance level is the probability of rejecting the null hypothesis when it is really true. The values most often seen are 0.05, 0.01, and 0.10.

3. Select the appropriate test statistic:

The choice of a test statistic depends on both the size of the sample and the characteristics of the data. Z-scores, t-scores, chi-square statistics, and F-statistics are often used test statistics.

4. Establish the Decision Rule: o Calculate the critical value(s) for the chosen significance level by examining the statistical distribution of the test statistic. The region(s) of rejection refers to the area(s) in a statistical hypothesis test where the null hypothesis is rejected in favor of the alternative hypothesis.

5. Data Collection and Analysis: Gather the sample data and calculate the test statistic.

6. Reach a decision: Compare the test statistic to the crucial value(s). Reject the null hypothesis if the value of the test statistic falls inside the predetermined range of rejection. On the other hand, the null hypothesis should not be dismissed.

7. Analyze the findings: To make inferences based on the choice. Rejecting the null hypothesis indicates that there is sufficient evidence to support the alternative hypothesis.

15.9 COMMONLY CONDUCTED TESTS

The Z-test is used when the population variance is known or when the sample size is big ($n > 30$).

The t-test is used when the population variance is unknown and the sample size is small ($n < 30$).

The one-sample t-test is a statistical test that compares the mean of a sample to a known value.

The two-sample t-test compares the means of two distinct groups that are not dependent on each other.

The paired t-test is a statistical test that compares the means of the same group at different points in time.

The chi-square test is used to assess the association between two variables or the goodness of fit in categorical data.

• ANOVA (Analysis of Variance): A statistical technique used to compare the means of three or more groups.

Above, we have outlined many significant tests often used to evaluate hypotheses, which serve as the foundation for making critical judgments. However, it is important for a researcher to constantly keep in mind the many limits of these experiments. The significant constraints are as follows:

(i) The tests should not be used in a rigid or automatic manner. It is important to remember that testing is not the act of making decisions; rather, tests serve as helpful tools for decision-making. Therefore, a correct understanding of statistical evidence is crucial for making informed judgments.

(ii) Tests do not provide explanations for the underlying causes for the difference in means between the two samples. These tests only determine whether the difference is caused by sample variations or other factors, but they do not identify the specific causes for the discrepancy.

(iii) The results of significance tests rely on probabilities and hence cannot be communicated with complete confidence. When a test

demonstrates statistical significance, it indicates that the observed difference is unlikely to be a result of random chance.

(iv) Statistical conclusions derived from significance tests cannot be considered completely accurate evidence about the validity of the hypotheses. This is particularly true when dealing with small samples, since the likelihood of making incorrect conclusions tends to be greater.

In order to enhance dependability, it is necessary to appropriately increase the size of samples.

These constraints indicate that in situations involving statistical significance, the inference procedures (or tests) should be used in conjunction with a thorough understanding of the subject matter and the capacity to exercise sound judgment.

15.10 LET US SUM UP:

Estimation theory is a specialized area within signal processing and statistics that focuses on determining parameter values based on empirical or measurable data. The primary goal of estimate theory is to deduce the values of one or more parameters based on a set of observable data. It is crucial in several fields, including as finance, economics, engineering, and the natural and social sciences.

Hypothesis testing is the use of data to evaluate a hypothesis about a population. A hypothesis is a statement that asserts a certain attribute of a population. As an example, the null hypothesis asserts that the mean value of the population is 5. A test statistic is a numerical metric used to evaluate the soundness of a hypothesis.

Hypothesis testing is a statistical method used to assess the validity of a statement or assumption about a parameter of a population.

15.11 KEY WORDS:

Statistical estimator: is a mathematical technique or formula that calculates an approximation of a population parameter using data obtained from a sample.

Hypothesis testing: is a statistical method used to draw conclusions or make assumptions about a population based on the analysis of sample data.

15.12 ANSWERS TO CHECK YOUR PROGRESS :

1. A ..... is a mathematical technique or formula that calculates an approximation of a population parameter using data obtained from a sample.

Answer: statistical estimator

2.The..... of an estimator refers to the disparity between the predicted value of the estimator and the actual value of the parameter being evaluated.

Answer: bias

3. An estimator is considered to be..... if its predicted value is identical to the real parameter value.

Answer: unbiased

4. The..... of an estimator refers to the extent of dispersion in the estimator's sample distribution.

Answer: variability

5. The..... is the most popular unbiased estimate for the population mean.

Answer: sample mean

8. The..... Hypothesis is the proposition being examined in a hypothesis test and is often represented as H0.

Answer: null

7. A .....estimate delivers a single number as an estimate of a population parameter.

Answer: point

15.13 TERMINAL QUESTIONS

1. Discuss about Estimation Theory**.**
2. Explain Point Estimate and Interval estimation.
3. What do you mean by Hypothesis? Explain types of hypothesis.
4. Explain Hypothesis testing procedures.

UNIT 16: T-TEST AND Z-TEST

Structure

16.0 Objectives

16.1 **t-test**

16.2 Characteristics of the t-distribution

16.3 Procedure for t-test

16.4 t- Test for two samples

16.5 Z- Test

16.6 Let Us Sum Up

16.7 Key Words

16.8 Answers to Check Your Progress

16.9 Terminal Questions

16.0 OBJECTIVES

After studying this unit, you should be able to:

- Acquire knowledge about the objective of the t-test and its appropriate use.
- Conduct a comparison of means from two independent samples to ascertain if they originate from the same population.
- Discover the objective of the Z-test and the circumstances in which it is used.
- Conduct a hypothesis test to see whether the average of a sample is substantially different from a known average of a population.

16.1 : T-TEST

Under the pseudonym "student," William Sealy Gosset made a substantial contribution to the development of significance tests for small samples. In 1908, he published a theoretical sampling distribution that is now commonly referred to as the "student's t-distribution." The symbol’t’ in the t-distribution denotes the ratio of the standard error of the sample mean to the difference between the sample mean and the population mean. In other words,—

$t = \frac{|\bar{X}-\mu|}{S}\sqrt{n}$

$\bar{X}$ represents the mean of the sample, μ represents the mean of the population, and S represents the most accurate estimate of the standard deviation of the population, as determined by the Properties of the t-Distribution ie. $\frac{\sum d^2}{n-1}$

16.2 CHARACTERISTICS OF THE T-DISTRIBUTION

The primary characteristics of the t-distribution are as follows:

(1) The t-distribution variable spans the entire range of minus infinity (∞) to plus infinity (∞), similar to a normal distribution.

(2) The normal distribution is less variable than the t-distribution. The t-Distribution approaches the normal distribution as n increases.

(3) The t-Distribution is symmetrical and mono-peaked, similar to the standard normal distribution. Nevertheless, the t-distribution curve is characterized by a more pronounced peak and extended tails.

(4) The t-curve's shape varies at different levels of significance, as evidenced by the t-table at the 5% and 1% levels of significance. Utilization of the t-distribution

16.3 PROCEDURE FOR T-TEST

The following procedure is implemented to employ the t-test to evaluate the significance of the difference between the mean of a random sample and the mean of the population:-

(1) Null hypothesis-Initially, this hypothesis is formulated to ensure that the population mean (u) is equivalent to the specified value of the mean (μ), i.e., H: μ=μ

The following is implied by this hypothesis: (a) The sample mean and the population mean are not significantly different, or

(b) the random sample is drawn from the normal population with a mean of μ.

(2) Calculation of the Test Statistic or t-statistic-The subsequent formula is employed for this purpose:-

$$t = \frac{|\bar{X}-\mu|}{S}\sqrt{n}$$

$\bar{X}$ Represents the mean of the sample, μ represents the mean of the population, n represents sample size and S represents the most accurate estimate of the standard deviation of the population

The standard deviation of the sample is determined using the following formula:

$$S = \sqrt{\frac{\sum(X-\bar{X})^2}{n-1}} \quad \text{or} \quad . \frac{\sum d^2}{n-1}$$

Note: 1. The formula will be modified as follows if deviations are taken from the presumed mean:

$$S = \sqrt{\frac{\sum d^2 x - \left(\frac{\sum dx}{n}\right)^2 X\, n}{n-1}}$$

Here dx represents the deviations from the presumed mean.

2. The formula for t will be altered as follows if the standard deviation of the sample ($\sqrt{\frac{\sum d^2}{n}}$)

is provided in the question:

$$T = \frac{|\bar{X} - \mu|}{S}\sqrt{n-1}$$

(3) Table value of t or critical value: This value is observed in the t-table at a specific level of significance and for degrees of freedom (n - 1) based on the queries.

(4) Decision-The null hypothesis is adopted if the calculated value of t is less than its tabulated or critical value. In other words, it has been demonstrated that there is no substantial disparity between the mean of the sample and the mean of the population. The hypothesis is rejected and the difference is deemed significant if the calculated value of t exceeds its table value.

The following are a few examples of the areas in which the t-distribution is typically employed to evaluate the significance of a variety of results obtained from limited samples:

Example:

The mean height of ten children randomly selected from a specific colony was 64 cm, with a standard deviation of 2.5 cm. Test the hypothesis that the average height of the children in the specified colony is less than 66 cm at a 5% level of significance. (The value of t for 9 d.f. at the 5% level of significance is 2.262.)

Solution:

In this question n= 10, Mean = 64 , Standard deviation= 2.5 and μ = 66

Null hypothesis: The children's average height is 66 cm. For instance, H0: μ = 66

Apply Formula for t value:

$$t = \frac{|\bar{X}-\mu|}{S}\sqrt{n-1}$$

$$= \frac{|64-66|}{2,5}\sqrt{10-1}$$

$$= \frac{|2|}{2,5}\sqrt{9}$$

$$= \frac{2x3}{2,5}$$

$$= \frac{6}{2,5}$$

$$= 2.4$$

Critical value: 2.262 (as indicated)

Decide: The hypothesis is rejected, as the average height of the children is 66 cms, as the calculated value of t (2.4) is more than its critical value (2.262).

Example:

Six students are randomly selected from a school and their Mathematics scores are 126, 126, 128, 132, 120, and 136 out of 200. Discuss the general observations that the mean marks in Hindi at the school were 132 in light of these marks.

Solution:

Null hypothesis: The average marks of the students are 132 cm. For instance, H0: μ = 132

Marks(X)	d from128	d^2
126	-2	4
126	-2	4
128	0	0
132	4	16
120	-8	64
136	8	64
∑X=768	∑d=0	$\sum d^2$=152

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

Mean = $\frac{768}{6}$

Mean = 128

Standard Deviation (S) = $\sqrt{\frac{\sum d^2}{n-1}}$

= $\sqrt{\frac{152}{6-1}}$

= $\sqrt{\frac{152}{5}}$

= $\sqrt{30.4}$

=5.51

Apply Formula for t value:

t = $\frac{|\bar{X}-\mu|}{S}\sqrt{n-1}$

= $\frac{|128-132|}{5.51}\sqrt{6}$

= $\frac{|4|}{2,5} x\ 2.44$

$= \frac{4x2.44}{5.51}$

$= \frac{9.76}{5.51}$

= 1.75

Critical value: n = 6, resulting in a d.f. = 6-1 =5. The critical value of 1 at the 5% level of significance and for a 5 d.f. is 2.571.

Decision: The null hypothesis is adopted because the calculated value of t = (1.75) is less than its critical value of (2.571). It implies that the school's average Hindi score was 132.

16.4 T- TEST FOR TWO SAMPLES

Testing the difference between the means of two tiny samples may have two objectives. (1) Have the two samples been obtained from the same population? or (2) Whether the factor that affects both samples is identical or if there is a substantial difference? The exam will be conducted in accordance with the following procedure:- (1) Null hypothesis Initially, the hypothesis is that the means of both samples are either identical or non-significantly different, i.e., H0: μ1 = μ2.

(2) Test statistic computation-The following formula is used to calculate the test statistic under the supposition that the population variances are unknown but equal ($\sigma 1^2 = \sigma 2^2 = \sigma 1^2$)

$$t = \frac{|\overline{X1} - \overline{X2}|}{S} \sqrt{\frac{n1n2}{n1+n2}}$$

$\overline{X1}$ and $\overline{X2}$ = Means of two sample 1 & 2, n_1 and n_2 Count of observations in two samples, S represents the sum of the standard deviations.

The value of S is determined in the following manner:

$$S=\sqrt{\frac{\sum(X1-\bar{X}1)^2+\sum(X2-\overline{X2})^2}{n1+n2-2}}$$

The aforementioned formula is employed exclusively when the dev iations (d) from the actual means are taken in both samples.

In the event that the actual means are in fractions, the deviations fr om the presumed means should be omitted for the sake of expedien cy. In this scenario, the formula will be as follows:

$$S=\sqrt{\frac{[\sum(X1-A1)^2-n1(\overline{X1}-A1)^2]+[\sum(X2-A2)^2-n2(\overline{X2}-A2)^2]}{n1+n2-2}}$$

X, and X2 = Actual values of the two samples , $\overline{X1}$ and $\overline{X2}$ = Actual means of two sample ,A, and A2 = Assumed averages of the two samples, n_1 and n_2 Count of observations in two samples,

S will be calculated as follows if the standard deviations of both sa mples are provided:

$$S=\sqrt{\frac{n1\sigma1^2+n2\sigma2^2}{n1+n2-2}}$$

(3) Degree of freedom: It is determined using the formula n + n2 + 2.

(4) Decision-A difference between the means of two samples is considered significant if the calculated value of t exceeds its critical or table value. Conversely, if the calculated value is less, the null hypothesis is adopted, and the difference is not deemed significant.

Example:

The following marks were acquired by two classes of students who participated in a test examination:

Class1	36	40	72	100	98	72	68	98	82
Class2	58	56	52	70	60	88	92		

Examine the significance of difference between mean marks secured by the above two classes.

Solution:

Null hypothesis: There is no significant difference between mean marks secured by two classes: H0: μ1 = μ2

Class 1			Class 2		
X1	d1	d_1^2	X2	d2	d_2^2
36	-38	1444	58	-10	100
40	-34	1156	56	-12	144
72	-2	4	52	-16	256
100	26	676	70	2	4
98	24	576	60	-8	64
72	-2	4	88	20	400
68	-6	36	92	24	576
98	24	576			
82	8	64			
∑X1=666	∑d1=0	$\sum d_1^2$=4536	∑X2=476	∑d2=0	$\sum d_2^2$=1544

Mean = $\frac{\sum X1}{n1}$

= $\frac{666}{9}$

= 74

Mean = $\frac{\sum X2}{n2}$

= $\frac{476}{9}$

= 68

$$S=\sqrt{\frac{\sum(X1-\bar{X}1)^2+\sum(X2-\overline{X2})^2}{n1+n2-2}}$$

$$S=\sqrt{\frac{4536+1544}{9+7-2}}$$

$$S=\sqrt{\frac{6080}{14}}$$

$$S=\sqrt{434.28}$$

S=20.83

$$t=\frac{74-68}{20.83}\sqrt{\frac{9x7}{9+7}}$$

$$t=\frac{74-68}{20.83}\sqrt{\frac{63}{16}}$$

$$t=\frac{74-68}{20.83}\sqrt{3.93}$$

$$t=\frac{14}{20.83}\sqrt{3.93}$$

$$t=\frac{14\ x\ 1.98}{20.83}$$

$$t=\frac{27.72}{20.83}$$

t=1.386

Degree of freedom: n1+ n2-2=9+7-2=14

Decision: The computed value of t is 1.386, and its table value at the 5% level of significance and for 14 d.f. is 2-145. Consequently, the null hypothesis is adopted, and there is no substantial difference in the mean scores of the two classes.

The mean lifespan of a sample of ten tube lights was determined to be 1,456 hours, with a standard deviation of 423 hours. A second sample of 17 tube lights selected from a distinct batch exhibited a mean life of 1,280 hours and a standard deviation of 398 hours. Is the mean of two samples significantly different?

Solution:

Given: n = 10, n = 17, X1 = 1,456, X2 = 1,280, σ = 423, σ2 = 398

Null hypothesis: There is no significant difference between mean of two samples: H0: μ1 = μ2

$\sigma 1^2 = 423^2 = 178929$

$\sigma 2^2 = 398^2 = 158404$

$S = \sqrt{\frac{10x178929+17x158404}{10+17-2}}$

$S = \sqrt{\frac{1789290+2692868}{25}}$

$S = \sqrt{\frac{4482158}{25}}$

$S = \sqrt{179286.32}$

S= 423.42

$t = \frac{1456-1280}{423.42}\sqrt{\frac{10x17}{10+17}}$

$t = \frac{176}{423.42}\sqrt{\frac{170}{27}}$

$t = \frac{176}{423.42}\sqrt{6.29}$

$t = \frac{176\ x\ 2.50}{423.42}$

$t = \frac{440}{423.42}$

t=1.04

Degree of freedom: n1+ n2-2=10+17-2=25

Decision: The critical value of t is 2.06, while the computed value is 1.04. Therefore, the null hypothesis is valid, and there is no substantial difference in the mean values of the two samples.

Example:

The following marks were acquired by two classes of students who participated in a test examination:

Class 1	48	32	40	44	30	14	20	34
Class 2	31	18	26	45	22	28	40	

Examine the significance of difference between mean marks secured by the above two classes.

Solution:

Null hypothesis: There is no significant difference between mean marks secured by two classes: H0: $\mu 1 = \mu 2$

Class 1			Class 2		
X1	(X_1-A_1) A1=33	(X1-A1)2	X2	(X2-A2) A1=30	$(X2-A2)^2$
48	15	225	31	1	1
32	-1	1	18	-12	144
40	7	49	26	-4	16
44	11	121	45	15	225
30	-3	9	22	-8	64
14	-19	361	28	-2	4
20	-13	169	40	10	100
34	1	1			

∑X=262	∑(X1-A1)=-2	∑(X1-A1)2=936	∑X2=210	∑(X2-A2)=0	∑(X2-A2)2=554

Mean = $\frac{\sum X1}{n1}$

= $\frac{262}{8}$

= 32.75

Mean = $\frac{\sum X2}{n2}$

= $\frac{210}{7}$

= 30

S= $\sqrt{\frac{[\sum(X1-A1)^2-n1(\overline{X1}-A1)^2]+[\sum(X2-A2)^2-n2(\overline{X2}-A2)^2]}{n1+n2-2}}$

S= $\sqrt{\frac{[936-8(32.75-33)^2]+[554-7(30-30)^2]}{8+7-2}}$

S= $\sqrt{\frac{[936-8x0.0625]+[554-7x0]}{8+7-2}}$

S= $\sqrt{\frac{[936-0.5]+[554]}{13}}$

S= $\sqrt{\frac{1489.5}{13}}$

S= $\sqrt{114.58}$

S= 10.7

t = $\frac{132.75-30}{10.7}\sqrt{\frac{8x7}{8+7}}$

t = $\frac{2.75}{10.7}\sqrt{\frac{56}{15}}$

$t = \frac{2.75}{10.7}\sqrt{3.73}$

$t = \frac{2.75\ x\ 1.93}{10.7}$

$t = \frac{5.31}{10.7}$

t=0.497

Degree of freedom: n1+ n2 -2 = 8+7-2=13

Decision: The computed value of t is 0.497, and its table value at the 5% level of significance and for 13 d.f. is 2.16. Consequently, the null hypothesis is adopted, and there is no substantial difference in the mean scores of the two classes.

16.5 Z- TEST

A Z test is a statistical test that is employed to ascertain whether there is a substantial difference between the means of two groups. It is generally employed when the population variance is known and the sample size is large (usually n > 30).

The primary procedures for conducting a Z test are as follows:

Z-Test for a Single Sample A one-sample z-

test is used to assess whether there is a difference between the sam ple mean and the population mean when the population standard de viation is known.

The population mean is represented by the symbol μ, the sample m ean is represented by $\bar{x}$, the population standard deviation is repres ented by σ, and the sample size is indicated by n.
The procedure for doing a one-sample z-test using the z-
test statistic is as follows:
The null hypothesis is stated as Ho: μ = μo. Left-Tailed Test:

Alternate Hypothesis: H1: μ < Ho Decision Criteria: Reject the nul l hypothesis if the z statistic is less than the z critical value.
The null hypothesis is stated as Ho: μ = μo. Right-Tailed Test: Alternative Hypothesis: H1: H>

Hypothesis Decision Criteria: Reject the null hypothesis if the z sta tistic is greater than the z critical value.

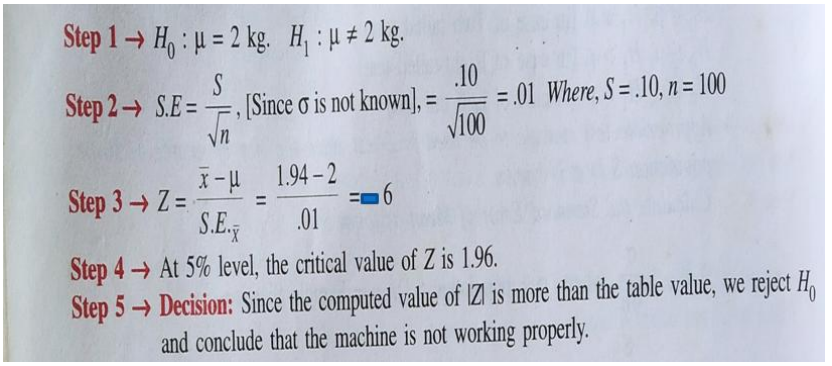
Null Hypothesis: Ho: μ = Ho Two-Tailed Test:

Alternate Hypothesis: H1: μ ≠ Ho Decision Criteria: Reject the nul l hypothesis if the z statistic exceeds the z critical threshold.
Two samples for Z-tests The purpose of a two-sample z-test is to assess if there is a difference between the means of two sa mples.
The first sample's sample mean, population mean, and population v ariance are represented by $(x1)^{-}$, μ1, and σ1 respectively.The secon d sample's sample mean, population mean, and population variance are represented by $(x2)^{-}$, μ2, and σ2, respectively.

Example:

The fertilizer mixing machine is programmed to dispense 4 kilograms of nitrate for every 100 kilograms of fertilizer. Five bags weighing 100 kilograms each are inspected. The percentage of nitrate is 2, 6, 4, 3, and 1.Is there evidence to suggest that the machine is faulty?

Solution:



16.6 LET US SUM UP:

William Sealy Gosset, under the alias "student," made a significant contribution to the advancement of significance tests for small sample sizes. In 1908, he introduced a theoretical sampling distribution that is well known today as the "student's t-distribution." The symbol 't' in the t-distribution represents the ratio of the standard error of the sample mean to the difference between the sample mean and the population mean.

A Z test is a statistical test used to determine whether there is a significant difference between the means of two groups. It is often used when the population variation is known and the sample size is big (usually n > 30).

16.7 KEY WORDS:

Z- test: is a statistical test that is employed to ascertain whether there is a substantial difference between the means of two groups.

t-test : is a statistical test used to ascertain if there exists a notable disparity between the means of two groups.

16.8 ANSWERS TO CHECK YOUR PROGRESS:

1. A t-test is performed when the sample size is .....

Answer: Small

2. In a t-test, the distribution that is utilized to establish the critical value is the _____ distribution.

Answer: t

3. In a Z-test, the .....distribution is employed to find the critical value.

Answer: normal

4. The t-test is used to compare the means of ..... groups.

Answer: one or two

5. In a Z-test, the ..... standard deviation is employed.

Answer: population

6. The degrees of freedom in case of t-test is .....

Answer: n-1

7. The ..... value for a Z-test is calculated from the Z-table.

Answer: critical

16.9 TERMINAL QUESTIONS:

1. What do you mean by t-test ? Explain procedure for calculation of t-test.

2. What do you mean by Z-test ? Explain procedure for calculation of Z-test.

3. Ten students are randomly selected from a school and their Mathematics scores are 63, 61, 64, 66, 60, 68, 67, 68, 70 and 71 out of 100. Discuss the general observations that the mean marks in Hindi at the school were 68 in light of these marks.

4. The following marks were acquired by two classes of students who participated in a test examination:

Class1	50	29	41	44	30	31	20	21
Class2	41	30	26	34	22	28	31	

Examine the significance of difference between mean marks secured by the above two classes.

5. Ten individuals are chosen at random from a population and their heights are found to be in inches, 63, 63, 64,65,66,69,69,70,70 and 71. In the light of these data, discuss the suggestion that the mean height in the universe is 65 inches. (Table value of t at 5% level of significance for 9 d.f. is 2.26)

UNIT 17: F-TEST AND ANOVA

Structure

17.0 Objectives

17.1 F-test

17.2 Assumptions for F-test

17.3 Procedure for calculating the F-test

17.4 Analysis of Variance (ANOVA)

17.5 Assumptions for analysis of variance

17.6 Applications of analysis of variance

17.7 Method for analyzing variance

17.8 Let Us Sum Up

17.9 Key Words

17.10 Answers to Check Your Progress

17.11 Terminal Questions

17.0 OBJECTIVES

After studying this unit, you should be able to:

- Comprehend the objective and fundamental principles of the F-test and ANOVA.
- Familiarize yourself with the F-test, a statistical method used to assess the differences in variances between two populations or groups.
- Acquire knowledge of ANOVA, since it is used to compare variances and means among many groups.
- Utilize two-way ANOVA to analyze the impact of two independent factors on a dependent variable, as well as their interaction effects.

17.1: F-TEST

The F-Test, often known as Fisher's F-Test or Variance Ratio Test

The F-test is sometimes referred to as Fisher's F-test or the Variance Ratio test. The concept was first proposed by R.A. Fisher and then expanded upon by G. W. Snedecor. The F-test is named after Fisher as a tribute to his contributions.

The F-test is a statistical test that is used to determine the significance of the difference between the variances of two or more groups or populations. It is often used in hypothesis testing to assess if the means of the groups are significantly different from each other.

The F-test is a hypothesis test that compares the variances of two samples. This test is specifically used to determine the importance of the difference between two variances. It helps determine if two samples can be considered as coming from the same normal population with equal variances or not. The variance ratio test is named as such because it involves the calculation of the ratio of variances.

17.2 ASSUMPTIONS FOR F-TEST

The F-test relies on the following assumptions:-

(1) Normality: The population from which samples are taken follows a normal distribution.

(2) The random technique and independence - The items in the samples have been chosen in a random and independent manner.

(3) The variance ratio must be equal to or larger than 1. In this situation, the bigger estimate of variance is divided by the smaller estimate of variance.

(4) The additive property states that the total variance is equal to the sum of the variation between samples and the variance within samples.

Computing the F-test or hypothesis test for the variance of two populations:

17.3 PROCEDURE FOR CALCULATING THE F-TEST

The procedure for calculating the F-test is as follows:

Firstly, the variances of each samples are determined using the following formulas:

$$\mathbf{S_1^2 = \sum(X_1 - \overline{X1})^2/n_1\text{-}1}$$

$$\mathbf{S_2^2 = \sum(X_2 - \overline{X2})^2/n_2\text{-}1}$$

If the variances of both samples are provided in the question, the following formulas are used to apply Bessel's correction:

$$\mathbf{S_1^2 = n_1 \sigma_1^2/n_1\text{-}1}$$

$$\mathbf{S_2^2 = n_2 \sigma_2^2/n_2\text{-}1}$$

(2) The null hypothesis is stated, which might be one of the following two: (a) Two samples are randomly selected from the same population, or (b) The variances of the populations corresponding to both samples are identical, denoted as H0 = S_1^2 = S_2^2.

3. The variance ratio, often known as the F-statistic, is calculated in the following manner:-

F = Larger Estimate of Variance/Smaller Estimate of Variance

It means that if $S_1^2 > S_2^2$ the formula will be F= S_1^2 / S_2^2 and if $S_2^2 > S_1^2$ the formula will be F= S_2^2 / S_1^2

(4) Calculation of degrees of freedom - The degrees of freedom for a sample with a bigger variance(v1) and the degrees of freedom for a sample with a lower variance (v2).

(5) F-Table value: The crucial value of F may be calculated from the F-table at either the 5% or 1% level of significance for V1 and V2.

(6) Interpretation: The calculated value and tabulated value of F are compared. If the computed value of F is less than the tabulated value, the variance ratio will be regarded unimportant. In this case, both samples will be assumed to have been chosen from the same universe or from universes with the same variances. If the value of F is greater than the critical value of F, the null hypothesis will be rejected and the variance ratio will be deemed statistically significant.

Example:

The data provided pertains to a randomly selected sample of gover nment workers in two states of the Indian union.
Conduct a hypothesis test to see whether the variances of the two p opulations are equal.

	Group I	**Group 2**
Sample size	16	25
Mean monthly income (Rs.)	880	970
Sample variance	80	84

Solution:

The null hypothesis states that the variances of two populations are equal, or $S_1^2 = S_2^2$. The given information is as follows:

n1 = 16, X_1 = 880, $_1^2$ = 80 and n2 = 25, X_2 = 460, σ_2^2= 84

Following Bessel's adjustment:

S_1^2 = n_1 σ_1^2/n_1-1

= 16 x 80/16 -1

= 1280/15

=85.33

S_2^2 = n_2 σ_2^2/n_2-1

= 25 x 84/25 -1

= 2100/24

=87.53

F = Larger Estimate of Variance/Smaller Estimate of Variance

F= S_2^2 / S_1^2

= 87.53 / 85.33

= 1.026

v1 = 25-1 =24 and v2 = 16-1 = 15

The table value of F at a significance level of 5% for degrees of freedom v1=24 and v2=15 is 2.29.

Therefore, the F statistic with a value of 1.026 is less than the table value (2.29) of F. Therefore, the null hypothesis is accepted, indicating that the variances of both populations are identical.

Example

The sum of squares of deviations from the mean for two independe nt samples, with sizes 9 and 8 respectively, were found to be 160 a nd 91.

Are the samples considered to be selected from normal populations with equal variance?

Given $F_{0.05}(8, 7) = 3.73$

Solution:

The null hypothesis states that the variances of two populations are equal, or $S_1^2 = S_2^2$.

$$\mathbf{S_1^2 = \sum(X_1 - \overline{X1})^2/n_1\text{-}1}$$

$$\mathbf{= 160/9\text{-}1}$$

$$\mathbf{= 160/8}$$

$$\mathbf{= 20}$$

$$\mathbf{S_2^2 = \sum(X_2 - \overline{X2})^2/n_2\text{-}1}$$

$$\mathbf{= 91/8\text{-}1}$$

$$\mathbf{= 91/7}$$

$$\mathbf{= 13}$$

F = Larger Estimate of Variance/Smaller Estimate of Variance

$$\mathbf{F = S_1^2 / S_2^2}$$

$$= 20/ 13$$

$$= 1.54$$

The table value of F at a significance level of 5% for degrees of freedom $v_1=8$ and $v_2=7$ is $F_{0.05}(8, 7) = 3.73$.

Therefore, the F statistic with a value of 1.54 is less than the table value (3.73) of F. Therefore, the null hypothesis is accepted, indicating that the variances of both populations are identical.

Example:

A random sample of 10 dogs were given diet A. The weight gains, measured in pounds, within a specified time were as follows: 10, 6, 16, 17, 13, 12, 8, 14, 15, 9.

The weight gains in pounds for another random sample of 12 doss fed on diet B throughout the same time were as follows: 7, 13, 22, 15, 12, 14, 18, 8, 21, 23, 10, 17.

Conduct a test to determine whether both samples are derived from populations with equal variances. Given $F_{0.05}$ (11,9) = 3.112).

Solution:

The null hypothesis states that the variances of two populations are equal, or $S_1^2 = S_2^2$.

Sample X1			Sample X2		
X1	(X1-A1) A1=12	(X1-A1)^2	X2	(X2-A2) A1=15	(X2-A2)^2
10	-2	4	7	-8	64
6	-6	36	13	-2	4
16	4	16	22	7	49
17	5	25	15	0	0
13	1	1	12	-3	9
12	0	0	14	-1	1
8	-4	16	18	3	9
14	2	4	8	-7	49
15	3	9	21	6	36
9	-3	9	23	8	64
			10	-5	25
			17	2	4

∑X=120	∑(X1 - A1)= 0	∑(X1-A1)²=120	∑X2=180	∑(X2-A2)=0	∑(X2-A2)²=314

Mean = $\frac{\sum X1}{n1}$

= 120 / 10

=12

Mean = $\frac{\sum X2}{n2}$

= 180 / 12

= 15

S_1^2 = ∑(X_1 - $\overline{X1}$)²/n_1-1

= 120 / 10-1

= 120 / 9

= 13.33

S_2^2 = ∑(X_2 - $\overline{X2}$)²/n_2-1

= 314 / 12-1

= 314 / 11

= 28.55

F = Larger Estimate of Variance/Smaller Estimate of Variance

F= S_2^2 / S_1^2

= 28.55 / 13.33

= 2.142

The table value of F at a significance level of 5% for degrees of freedom v_1=11 and v_2=9 is $F_{0.05}$ (11,9) = 3.112.

Therefore, the F statistic with a value of 2.142 is less than the table value (3.112) of F. Therefore, the null hypothesis is accepted, indicating that the variances of both populations are identical.

17.4 ANALYSIS OF VARIANCE (ANOVA)

The t-test and x-test are very valuable for assessing the significance of the mean of a single sample or the significance of the difference in means between two samples. Nevertheless, if there are more than two samples, then these approaches are unsuitable. The issue is solved using the approach of analysis of variance, which was first introduced by Ronald A. Fisher in 1923 and then expanded upon by George W. Snedecor. The primary goal of analysis of variance is to assess the statistical significance of means from several samples using a single test. Analysis of Variance refers to the process of statistically examining the differences between groups or treatments in order to determine whether there are any significant variations. It involves comparing the variances within and between groups to assess the impact of different factors on the observed data.

The analysis of variance is a mathematical procedure that breaks down the homogeneity of variation into different components of variance. Sir Ronald A. Fisher's definition of analysis of variance is the process of distinguishing the variation attributable to one set of causes from the variance attributable to other groups. Owen L. Davies states that the analysis of variance is the fundamental process of identifying components that correspond to the origins of variances.

Yule and Kendall state that the analysis of variance is mostly used to assess for homogeneity among distinct sets of data.

Analysis of variance is a method used to partition the overall variation into variations generated by distinct causes and to assess the homogeneity of different sample means. The statistical method known as analysis of variance is often abbreviated as 'ANOVA'. Elements of overall variance-

Typically, the overall variance is divided into two components:

There are two types of variance:

(1) variance between samples and

(2) variance within samples. The total variance is equal to the variance between samples plus the variance within samples.

The intergroup variance quantifies the variation between distinct samples, whereas the intragroup variance represents the overall variance inside each individual sample.

Assumptions of Analysis of Variance include some conditions that need to be met in order to ensure the validity and accuracy of the analysis.

17.5 ASSUMPTIONS FOR ANALYSIS OF VARIANCE

The analysis of variance approach is based on the following key assumptions:

(1) Normality: The population from which the different samples have been chosen follows a normal distribution. Furthermore, each of the samples is a random sample selected without bias or preference.

(2) Independence- Each sample is mutually exclusive and unrelated to the other samples. If the samples are not independent, the analysis of variance may not be informative due to the presence of correlation.

(3) The additive property states that the sum of the variances of different components must be equal to the overall variance.

The analysis of variance (ANOVA) is a statistical technique that is used to determine the significance of differences between groups or treatments in an experiment. It is widely used in many fields such as psychology, biology, economics, and engineering. ANOVA allows researchers to assess the impact of several factors on a dependent variable and identify any significant relationships. This analysis is crucial in hypothesis testing, experimental design, and decision-making processes, as it provides valuable insights into the effects of various variables

Analysis of variance is a statistical approach that is used in almost all types of experimental designs. The tool is a crucial and potent instrument for doing research, as it enhances the conciseness and precision of the outcomes. According to Morrid Budin, systematic techniques for evaluating the extent of variation and identifying its origins are a valuable addition to statistical methodology.

17.6 APPLICATIONS OF ANALYSIS OF VARIANCE

The primary domains where analysis of variance is used include:

(1) To assess the statistical significance of variations among the means of many samplesThe analysis of variance approach is often used to assess the significance of variations between means of many samples. This test aids in determining if all samples have been

obtained from the same population or not, and whether the observed discrepancies are a result of sampling fluctuations or any other cause.

(2) To assess the statistical significance of variations between variances. The analysis of variance is also used to assess the importance of variations between variances of distinct samples. In this context, the F-coefficient or variance ratio is computed.

(3) Two-way classification is a useful method for analyzing data when it is categorized into different groups based on two variables. By using the methodology of analysis of variance, we may make significant conclusions about the similarity or homogeneity within these categories. For instance, a company entity hires four sales representatives and collects sales data for each season individually, namely Summer, Winter, and Rainy seasons. Subsequently, the significance of the disparity may be assessed by a two-way analysis of variance.

(4) The analysis of variance approach is used to examine the linear character of regression, the significance of the coefficient of correlation, and the relevance of multiple correlation in order to assess the significance of correlation and regression.

17.7 METHOD FOR ANALYZING VARIANCE:

1.One-way classification

The method used for analyzing variance in a one-way classification is called Analysis of Variance (ANOVA).

When data are categorized based on a single criteria, it is referred to as one-way or one-factor or one-fold classification. For instance, if

we gather data on agricultural productivity based on the use of various types of fertilizers, it would be classified as a single-factor classification as all other variables have been disregarded except for the fertilizer used.

One - Way ANOVA

1. Null Hypothesis- First of all, this hypothesis is formulated that means of populations from k samples have been randomly drawn are equal to one another, i.e. There is no significance difference between these means:-

$H0 = \mu1 = \mu2 = \mu3 = \ldots\ldots \mu k$

2. Total of samples items:- first of all total of all items of each samples.

$T = \sum X_1 + \sum X_2 + \sum X_3 + \ldots\ldots\ldots \sum X_k$

3.Total sum of Squares of all items:- Each value of each sample is squared and find the total of all squares of each sample.

$\sum X_1^2, \sum X_2^2, \sum X_3^2, \ldots\ldots\ldots \sum X^2_k$

4. Correction factor:-

$C.F. = T^2/N$

5. Total sum of squares or SST:
$SST = [\sum X^2_1 + \sum X^2_2 + \sum X^2_3 + \ldots\ldots\ldots \sum X^2_k] - T^2/N$

6. Sum of squares between samples or SSB

$SSB = [(\sum X_1)^2/n_1 + (\sum X_2)^2/n_2 + (\sum X_3)^2/n_3 + \ldots\ldots\ldots (\sum X_k)^2/n_k] - T^2/N$

7. Sum of squares within samples or SSW:-

SSW= SST – SSB

8. Analysis of Variance table and interpretation:-

ANOVA table

Source of Variation	**Sum of squares**	**Degree of freedom(d.f)**	**Mean squares**
1. Between Samples	SSB	V1= k-1	MSB
			#NAME?
2.Within samples	SSW	V2=N-k	MSW
			#NAME?

. 9. Calculation of F= MSB/ MSW If MSB > MSW or

F= MSW/ MSB If MSW > MSB

10.Interpretation-

If calculated value of F is less than table value of F than Null hypothesis is accepted.

or

If calculated value of F is more than table value of F than Null hypothesis is rejected.

Example:

The following table gives the yield of four plots each of four varieties of rice. Find out that the variety differences are significant or not.

The table value of F for 5% level of significance is 5.95.

Varieties of Rice (Yield in Kg.)			
A	**B**	**C**	**D**
8	**10**	**16**	**14**
10	**11**	**12**	**10**
10	**8**	**14**	**12**
8	**11**	**6**	**16**

Solution:

Null hypothesis:- There is no significance difference in four varieties of the rice.

Variety A		Variety B		Variety C		Variety D	
X_1	X_1^2	X_2	X_2^2	X_3	X_3^2	X_4	X_4^2
8	64	10	100	16	256	14	196
10	100	11	121	12	144	10	100
10	100	8	64	14	196	12	144
8	64	11	121	6	36	16	256
$\sum X_1$ = 36	$\sum X^2_1$ =328	$\sum X_2$ = 40	$\sum X^2_2$ = 406	$\sum X_3$ = 48	$\sum X^2_3$ =632	$\sum X_4$ =52	$\sum X^2_4$ =696

$T = \sum X1+\sum X2+\sum X3+\sum X4$

$= 36+40+48+52$

$=176$

Correction factor (C.F.) = T2 / N

$= (176)^2/16$

$= 30976 / 16$

$= 1936$

$SST= [\sum X^2_1+\sum X^2_2+ \sum X^2_3+ \ldots\ldots\ldots \sum X^2_k] - T^2/N$

$= (328+406+632+696) – 1936$

$=2062 - 1936$

$= 126$

$SSB= [(\sum X_1)^2/n_1+(\sum X_2)^2/n_2+ (\sum X_3)^2/n_3+ \ldots\ldots\ldots (\sum X_k)^2/n_k] -T^2/N$

$= [(36)^2/4+(40)^2/4+ (48)^2/4+ (52)^2/4] -1936$

$= (1296/4 + 1600/4 + 2304/4 +2704/4) – 1936$

$= (324 + 400 + 576 + 676) =1936$

$= 1976 – 1936$

$= 40$

ANOVA TABLE

Source of Variation	**Sum of squares**	**Degree of freedom(d.f)**	**Mean squares**	**Variance ratio**
1. Between Samples	**SSB=40**	**V1 = k-1=4-1=3**	**MSB**	**F= MSB/MSW = 13.33/7.17= 1.86**

2.Within samples	**SSW= 126-40 = 86**	**V2=N-k= 16-4=12**	**MSW =SSW/N-k = 86/12=7.17**	

Interpretation:

The table value is of F at 5% level of significance is 5.95 is greater than it's calculated value so Null Hypothesis is accepted. Hence differences in the varieties of rice are not significant.

2. Two Way - ANOVA

1. Null Hypothesis- First of all, this hypothesis relating to both factors are formulated .
H0

2. Coding of data if needed

3. Total of samples items:- first of all total of all items of each samples.

$T = \sum X_1 + \sum X_2 + \sum X_3 + \ldots\ldots \sum X_k$

K= number of samples

4.Total sum of Squares of all items:- Each value of each sample is squared and find the total of all squares of each sample.

$\sum X_1^2, \sum X_2^2, \sum X_3^2, \ldots\ldots \sum X^2_k$

5. Correction factor:-

C.F. = T^2/N, N =Total items in all samples

5. Total sum of squares or SST:

$SST = [\sum X^2_1 + \sum X^2_2 + \sum X^2_3 + \ldots\ldots\ldots \sum X^2_k] - T^2/N$

6. Sum of squares between columns or SSC

$SSC = [(\sum Xc_1)^2/nc_1 + (\sum Xc_2)^2/nc_2 + (\sum Xc_3)^2/nc_3 + \ldots\ldots (\sum Xc_k)^2/nc_k] - T^2/N$

7. Sum of squares between Rows or SSR

$SSR = [(\sum Xr_1)^2/nr_1 + (\sum Xr_2)^2/nr_2 + (\sum Xr_3)^2/nr_3 + \ldots\ldots (\sum Xr_k)^2/nr_k] - T^2/N$

8. Sum of squares due to error or residual sum of squares or SSE:-

SSE= SST – (SSC +SSR)

9. Analysis of Variance table and interpretation:-

Source of Variation	Sum of squares	Degree of freedom(d.f)	Mean squares	Variance Ratio
1. Between Samples	SSC	c-1	MSC =SSC/c-1	F= MSC/MSE
2. Between Rows	SSR	r-1	MSR =SSR/r-1	F= MSR/MSE
3. Residual	SSE	(c-1)(r-1)	MSE=SSE/ (c-1)(r-1)	

10. Interpretation-

If calculated value of F is less than table value of F than Null hypothesis is accepted.

or

If calculated value of F is more than table value of F than Null hypothesis is rejected.

A factory manager, seeking to purchase machines for a certain ope ration in the production process, acquired one machine from each o f the four businesses specializing in manufacturing such machines. The manager then assigned three workers, each of whom worked o ne day on each of the four machines, in a random sequence.
Below are the units that are obtained as a consequence.

Workmen→ Machine↓	W1	W2	W3
M1	92	93	94
M2	94	96	98
M3	97	97	100
M4	98	99	99

Discuss the significance of variation of production among the different type of machines and also among the workers. The table value of F for 1% level of significance is F(2,6)10.92 and F(2,6)9.78 respectively.

H0 : (i)The workers do not differ with respect to mean productivity (columns W)

(ii) The mean production does not differ for four machines (rows M)

Coded Data X= 96					Squares			
Workmen→	W_1	W_2	W_3	$\sum X_r$	W^2_1	W^2_2	W^2_3	$\sum X_r$

Machine ↓								
M1	-4	-3	-2	-9	16	9	4	29
M2	-2	0	2	0	4	0	4	8
M3	1	1	4	6	1	1	16	18
M4	2	3	3	8	4	9	9	22
$\sum X_C$	-3	1	7	T= +5	25	19	33	77

Analysis of Variance table and interpretation

Source of Variation	Sum of squares	Degree of freedom(d.f)	Mean squares	Variance Ratio
1. Between Samples	SSC= 12.667	2	MSC 6.33	F= MSC /MSE = 9.45
2. Between Rows	SSR= 58.250	3	MSR 19.42	F= MSR /MSE= 29
3 Residual	SSE= 4	6	MSE= 0.67	

Interpretation:

1. The calculated value of F is 9.45, while table value of F at 1% of level of significance is 10.92. Hence, hypothesis is acceptable in case of mean productivity of workers. In other words the workers do not differ in respect of mean productivity.

2. The calculated value of F is 29, while table value of F at 1% of level of significance is 9.78. Hence, hypothesis is rejected in case of

mean production of machines. In other words the machines are differ in respect of mean productivity.

17.8 LET US SUM UP:

The F-test is also known as Fisher's F-test or the Variance Ratio test. The idea was first introduced by R.A. Fisher and further elaborated by G. W. Snedecor. The F-test is named after Fisher as an homage to his significant achievements.

The F-test is a statistical test used to assess the significance of the variation between the variances of two or more groups or populations. It is often used in hypothesis testing to evaluate if the means of the groups exhibit significant differences.

Analysis of variance, first proposed by Ronald A. Fisher in 1923 and subsequently refined by George W. Snedecor. The main objective of analysis of variance is to evaluate the statistical significance of means from several samples using a single test. Analysis of Variance (ANOVA) is a statistical procedure used to compare and assess the differences across groups or treatments, with the aim of identifying any significant variances. The process entails evaluating the differences in variability both within and between groups in order to determine the influence of various variables on the collected data.

17.9 KEY WORDS:

ANOVA: Analysis of variance

SSC: Sum of squares between columns

SSB: Sum of squares between samples

17.10 ANSWERS TO CHECK YOUR PROGRESS:

1. ANOVA stands for Analysis of.....

Answer: Variance

2.The F-test is used to examine the ..... between group variances and within-group variances.

Answer: disparity

3.In ANOVA, the F-statistic is computed by dividing the ..... variance by the within-group variance.

Answer: between-group

4.The null hypothesis in ANOVA asserts that all group means are .....

Answer: equal

5. The degrees of freedom for the numerator in the F-test is determined as the number of groups minus .....

Answer: one

6. A ..... F-value in the F-test implies that the between-group variation is greater than the within-group variance.

Answer: larger

17.11 TERMINAL QUESTIONS:

1. Clarify what "analysis of variance" means. Analyze the variance for both one-way and two-way categories in a short description.

2. Enumerate the fundamental assumptions of the analysis of variance.

3. Explain Method for analyzing variances.

4. A general test conducted in Mangalayatan University of certain students of MBA selected at random from four branches. Marks

obtained individually are as follows. Carry out analysis of variance to test significance difference between mean marks of four branches.

Human Resource (A)	**Finance (B)**	**Marketing (C)**	**International Business (D)**
13	**8**	**4**	**6**
7	**8**	**6**	**7**
10	**8**	**4**	**6**
11	**6**	**6**	**8**
9	**5**	**5**	**3**

(Table Value of $F_{3,16}$ is 3.239, level of significance 5%.)

5. Explain procedure for calculation of F-test.

UNIT 18: CHI-SQUARE TEST ($\chi 2$)

Structure

18.0 Objectives

18.1 Introduction

18.2 Properties of the chi-square distribution

18.3 Applications of the Chi-Square Test

18.4 Requirements or criteria for the application of the x2 test

18.5 Method for calculating the Chi-Square (χ^2):

18.6 Constraints in the application of the x2 test

18.7 Let Us Sum Up

18.8 Key Words

18.9 Answers to Check Your Progress

18.10 Terminal Questions

18.0 OBJECTIVES

After studying this unit, you should be able to:

- Comprehend the Chi-Square test and its significance in statistical analysis, especially for data that is categorized into different groups.
- Acquire knowledge about the many categories of Chi-Square testing.
- Comprehend the Chi-Square Test of Goodness of Fit, which is used to assess the degree of conformity between an observed distribution and a predicted distribution.
- Comprehend the Chi-Square Test of Independence, used to evaluate the independence or association between two categorical variables.

18.1 INTRODUCTION:

Chi-square test The chi- square test, often known as the x2 test, is a very valuable statistical test in the field of statistical research. Karl Pearson initially used this test in 1900, and it has since become a commonly employed non-parametric test in statistical analysis due to its simplicity.

The Chi-Square test ($\chi 2$) is a significant and often used hypothesis test. It is essentially a data analysis that relies on observations of a randomly selected group of variables. Typically, it involves comparing two sets of statistical data. This method is applicable when we possess data that is represented by frequencies or falls within the categories of ordinal or nominal levels of measurement.

The chi- square test is a statistical technique that quantifies the discrepancy between observed frequencies (fo) and predicted frequencies (fe). Additionally, it clarifies if the disparity is substantial or just a result of random variations in the sample. Mathematically, it is represented by the following equation:

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

χ^2 = Value of Chi- Square

O = Observed frequencies

E = Expected frequencies.

18.2 PROPERTIES OF THE CHI-SQUARE DISTRIBUTION

The primary attributes of the Chi-square test are as follows:

(1) The Chi-square test is a non-parametric test. The x2 test relies on observed and predicted frequencies or occurrences, while other tests such as the Z test, t-test, and so on, rely on measures such as mean and standard deviation.

(2) Zero sum of differences between actual and expected frequencies-The total sum of the differences between the actual frequencies (fo) and the anticipated frequencies (fe) will always be zero, meaning that the sum of the actual frequencies (Σfo) minus the sum of the expected frequencies (Σfe) equals the total number of observations (N) minus the total number of observations (N), which is equal to zero.

(3) Uninterrupted dispersionThe -x2 distribution is a kind of continuous distribution. However, it may be used to discrete random variables that can be quantified and organized into tables, with or without grouping.

The Chi-square distribution is only determined by the degrees of freedom (d.f.), which is the only parameter used in the Chi-square test. The determination of the critical value of Chi-square is based on the supplied values in the question. Each degree of freedom has its own specific critical value, which may be seen at various levels of significance.

(5) The Chi-square test is a valuable tool for assessing hypotheses. However, it is not deemed valuable for estimating, unlike the Z and t-test.

(6) The values of a Chi-square distribution graph are as follows: the mean is equal to the number of degrees of freedom (X = d.f.); the mode is always the degrees of freedom minus two (z = d.f. - 2), although it cannot be less than zero; the variance is equal to twice the degrees of freedom (Variance = 2 d.f.). It is important to emphasize that Chi-square values cannot be negative.

The chi-square distribution and its degrees of freedom (d.f.) are denoted by (7).The Chi-square distribution's form is contingent upon the degrees of freedom. Varying degrees of freedom will result in distinct curve shapes. The x2 distribution exhibits positive right skewness when the degrees of freedom are low. As the number of degrees of freedom grows, the distribution gets less skewed and approaches symmetry more quickly.

18.3 APPLICATIONS OF THE CHI-SQUARE TEST

The Chi-square test has emerged as a very effective technique for evaluating the hypothesis of many statistical issues. The Chi-square test has many significant applications:The user did not provide any text.

(1) Independence test The Chi-square test is used to assess the correlation or lack thereof between two sets of characteristics. For instance, we may assess the efficacy of quinine in managing malaria, determine whether there is a correlation between the intellect levels of fathers and sons, or ascertain if they are unrelated. To determine the association between qualities, we calculate the anticipated frequencies assuming that the attributes are independent. We next calculate the Chi-square value. The computed value of x2 is compared to the critical value of chi-square from a table, based on the provided level of significance and degrees of freedom. If the computed chi-square value is less than or equal to the table value, it indicates that the characteristics are independent. Conversely, if the computed chi-square value exceeds the table value, it indicates that the qualities are related.

The chi-square test highlights the importance of not just the difference between the actual frequencies of the sample and the

predicted frequencies, but also the need for this difference to be both noticeable and statistically significant in order to infer that the same connection exists in the population as well.

(2) Goodness of fit test - The Chi-square test is used to assess if observed frequencies align with a particular theoretical frequency distribution, such as Binomial, Poisson, Normal, etc. whether there is a disparity between the observed frequencies and the theoretical frequencies, an analysis is conducted to determine whether this disparity is statistically significant or not. Essentially, the test of goodness of fit determines whether the sample data is consistent with the theoretical distribution or not. If the computed value of x2 exceeds the tabulated value, the adequacy of the fit is deemed unsatisfactory. Conversely, if the computed value of x2 is lower than the tabulated value, the fit is deemed satisfactory. This means that the discrepancy between the observed and predicted frequencies is ascribed to random variations caused by sampling.

(3) Homogeneity test-This test determines if two or more independent random samples have been chosen from the same population or not. The chi-square test is used to assess population variation.

18.4 REQUIREMENTS OR CRITERIA FOR THE APPLICATION OF THE X2 TEST

(1) Adequate sample size - The overall number of observations or frequencies employed in this test must be sufficiently big. While 'fairly big' is a relative concept, it is generally advised not to employ this test when the value of N is less than 50.

(2) The predicted frequency of any item or cell should not be considered modest if it is less than 5. If the value is less than 5, the frequencies from the neighboring items or cells are combined in

order to get a total of 5 or more. In this scenario, Yate's adjustment may also be used.

(3) Data in original units - The data should be presented using the original units of measurement, meaning they should not be represented as percentages or proportions.

(4) Random sampling - The data acquired for this test should be obtained using a random and unbiased selection process.

(5) The occurrences that this test is being applied to must be mutually exclusive.

(6) Linear constraints refer to limitations on cell frequencies that are expressed in a linear manner, meaning they do not use square or higher powers of frequencies. The cell frequencies must adhere to linear limitations, meaning they should not include squared or higher powers of the frequency. Specifically, the sum of observed frequencies ($\sum O$) should be equal to the sum of expected frequencies ($\sum E$), which in turn should be equal to N.

(7) The observed samples must be mutually independent.

The sample was chosen in a haphazard manner from the whole population. Every theoretical frequency should be significant.

18.5 METHOD FOR CALCULATING THE CHI-SQUARE (X^2):

The method for calculating the Chi-Square (χ^2) is as follows:

i. The investigation establishes a null hypothesis and an alternative hypothesis.

ii. A significance level is selected to determine if the null hypothesis should be rejected.

iii. A sample of observations is selected randomly from a statistically relevant population.

iv. The predicted or theoretical frequencies are calculated based on the supplied actual data using probability. Typically, this involves presuming that a certain probability distribution is relevant to the statistical population being studied.

Get ready Frequency table obtained from observation:

Observed frequency table (O)		
AB	aB	B
Ab	ab	b
A	a	N

Get ready anticipated frequency distribution

Expected frequency table (E)		
AB	aB	B
Ab	ab	b
A	a	N

Calculate expected frequency Values:

In this process, the frequencies that are actually seen are compared to the frequencies that are anticipated or predicted based on theory.

AB= $\frac{AxB}{N}$

aB= $\frac{axB}{N}$

Ab= $\frac{Axb}{N}$

ab= $\frac{axb}{N}$

Calculate the degree of freedom (d.f.) using the formula d.f. = (r-1)x(c-1), where r represents the number of rows and c represents the number of columns.

Calculate the value of the Chi-Square:

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

χ^2 = Value of Chi- Square

O = Observed frequencies

E = Expected frequencies.

Prepare Analysis table:

Value	**O**	**E**	**O-E**	**$(O-E)^2$**	**$(O-E)^2/E$**
AB					
aB					
Ab					
ab					
Value of Chi- square					**$\chi^2 = \sum(O-E)^2/E$**

- (a). Test of independence
- (b). Test of goodness of fit
- (c). Test of homogeneity

(a). Test of independence

This test is useful for identifying the correlation between two or more characteristics.

Method:

The process of organizing current occurrences or (O) The assumption of no relationship or difference

Computation of anticipated frequency (E)

Calculation of Chi-square ($\chi 2$)

Calculation of the number of independent variables in a system

Determining the chi-square table value

Evaluation of a hypothesis or conclusion

Example:

The following info is provided to you. Do these data substantiate the theory that there is a positive correlation between the intelligence of dads and their sons?

	Intelligent sons	Dull sons
Intelligent fathers	24	12
Dull fathers	32	32

Table value of chi square for 1 d.f. at 5% level of significance is 3.841.

Solution:

Null Hypothesis: There is no association between intelligence of fathers and sons.

Intelligent fathers =A, Dull fathers = a, Intelligent sons =B, Dull sons = b

Given data are:

AB=24, aB =32, Ab= 12, ab = 32

Observed Frequency table (O)		
(AB) 24	(aB) 32	(B) 56
(Ab) 12	(ab) 32	(b) 44
(A) 36	(a) 64	(N) 100

The independence factor E for the cell AB has been determined as AxB/N = 36x56/100 = 20.16 = 20. The remaining numbers may be derived by calculating the differences from the subtotals in the 9-square table.

Calculation of Chi- square (χ^2)

Class	O	E	O-E	$(O-E)^2$	$(O-E)^2/E$
AB	24	20	4	16	0.8
aB	32	36	-4	16	0.44
Ab	12	16	-4	16	1
ab	32	28	4	16	0.57
Value of Chi- square					$\chi^2 = \sum(O-E)^2/E=2.816$

Degree of freedom = (c-1) x (r-1) =(2-1) x (2-1) =1

Conclusion:

- The calculated value of χ^2 **(2.816)** is less than the table value of χ^2 at 5% level of significance for 1 d.f. (3.841). Hence, our null hypothesis is proved and the figures given in the question do not support the hypothesis that intelligent fathers have intelligent sons.

(b). Test of goodness of fit:

This approach is mostly used for assessing the adequacy of fit. It aims to determine whether there is a difference between an observed frequency distribution and an estimated frequency distribution.

Example:

You are provided with the count of books borrowed from a public library for a certain week, spanning six days. Conduct an experiment to see if there is a correlation between the quantity of books borrowed and the day of the week.

Days	Number of books borrowed
Monday	140
Tuesday	132
Wednesday	160
Thursday	148
Friday	134
Saturday	150

Examine Chi-square test. Table value of χ^2 for 5 d.f. At 5% level of significance is 11.07.

Solution:

- Null Hypothesis(H0): The number of books borrowed does not depend on the day of the week, i.e. equal number of books is borrowed everyday.
- Total number of books borrow in a week is = 864
- Expected frequency (E) on the basis of equal borrowing everyday = 864/6 = 144

Calculation of Chi- square (χ^2)

Days	O	E	O-E	(O-E)2	(O-E)2/E
Monday	140	144	-4	16	0.1
Tuesday	132	144	-12	144	1
Wednesday	160	144	16	256	1.8
Thursday	148	144	4	16	0.1
Friday	134	144	-10	100	0.7
Saturday	150	144	6	36	0.3
Value of Chi- square					∑(O-E)2/E=3.9

Conclusion:

The calculated value of χ^2 **(3.9)** is less than the table value of χ^2 at 5% level of significance for 1 d.f. (11.07). Hence, our null hypothesis is correct and the number of books borrowed does not depend on the day of the week.

(c). Test of homogeneity:

The chi-square test of homogeneity is a more advanced version of the chi-square test of independence. These tests determine if two or more independent samples are taken from the same population or from distinct populations.

Example:

A total of 200 packets of chips were randomly chosen from the production of each of the five machines. The quantity of faulty packs of chips discovered were 7, 11, 17, 13, and 7. Are there any notable disparities among the machines? The table value of the chi-square (χ2) statistic at a 5% level of significance for 4 degrees of freedom (d.f.) is 9.488.

Solution:

- Null Hypothesis(H0): There is no significant difference among the machines.
- As there are five machines, the total number of defective packets should be equally distributed among these machines.
- Expected frequency (E) on the basis of equal distribution among the machines = 55 /5 = 11
- Calculation of Chi- square (χ^2)

Machine	O	E	O-E	(O-E)2	(O-E)2/E
1	7	11	-4	16	1.5
2	11	11	0	0	0
3	17	11	6	36	3.3
4	13	11	2	4	0.4
5	7	11	-4	16	1.5
Value of Chi- square					∑(O-E)2/E=6.7

Conclusion:

The calculated value of χ^2 **(6.7)** is less than the table value of χ^2 at 5% level of significance for 1 d.f. (9.488.). Hence, our null hypothesis is correct and the difference among the five machines in respect of defective packets of chips is not significant.

18.6 CONSTRAINTS IN THE APPLICATION OF THE X2 TEST

While the Chi-square test is really a practical and helpful test, it is important to be aware of some caveats and restrictions to prevent any misapplication of the test:-

1. Adjustments for low frequencies-It is important to verify that the frequencies are not too low. Typically, it is recommended that each frequency be at least 10, but under no circumstances should it be less than 5. If there exists any frequency that is below 5, the frequencies are combined by combining adjacent classes or cells.

2.The omission of non-occurrence should be avoided - The frequencies of non-occurrence should be included in the computation of x2 test.

3.The x2 test requires that the data be presented in terms of actual frequencies. It implies that percentages, proportions, rates, and similar measures should be avoided.

4.The Chi-square test should not be employed when measurements have been repeated on the same set of units, since it would be redundant. For instance, the scores of 10 students before to receiving coaching and subsequent to receiving coaching. In this scenario, the chi-square test is unable to ascertain the effectiveness of coaching courses. Furthermore, if a certain group of individuals provides their viewpoint on many items, the Chi-square test cannot be used since a contingency table cannot be constructed using such data.

5.In order to ensure accurate testing, it is imperative to: (i) Accurately calculate the degrees of freedom, (ii) Properly identify the critical value of x2 from the x2 table, (iii) Define the null hypothesis in the correct form, (iv) Verify the various sub-totals and totals of actual and expected frequencies, and (v) Calculate the expected frequencies based on a rational basis.

18.7 LET US SUM UP:

Chi-square test The chi-square test, sometimes referred to as the x2 test, is a very important statistical test within the realm of statistical

research. Karl Pearson first used this test in 1900, and it has since gained widespread usage as a non-parametric test in statistical analysis owing to its straightforwardness.

The Chi-Square test ($\chi 2$) is a widely used and statistically significant hypothesis test. It is simply a data analysis method that depends on observations of a randomly chosen set of variables. Generally, it entails the comparison of two sets of statistical data. This strategy is suitable when we have data that is expressed as frequencies or belongs to the categories of ordinal or nominal measurement levels. The chi-square test is a statistical method that measures the difference between actual frequencies (fo) and expected frequencies (fe). Furthermore, it provides clarification on whether the discrepancy is significant or just a consequence of random fluctuations in the sample.

18.8 KEY WORDS:

χ^2 = Value of Chi- Square

O = Observed frequencies

E = Expected frequencies.

18.9 ANSWERS TO CHECK YOUR PROGRESS:

1. The Chi-square test is used to determine if there is a significant association between _____ variables.
Answer: categorical
2. The Chi-square statistic quantifies the disparity between expected and ..... frequencies.

Answer: actual

3. The Chi-square test is used to assess the ..... of variables in a contingency table.

Answer: independence

4. The Chi-square test for goodness-of-fit evaluates whether the observed distribution of _____ deviates from a theoretical distribution.

Answer: frequencies

5. The Chi-square test for goodness-of-fit is used to compare observed frequencies with _____ frequencies.

Answer: expected

18.10 TERMINAL QUESTIONS

1 What is the definition of Chi-square test? Elucidate its importance in statistical analysis.

2. What are the conditions for applying Chi Square test?

3. Explain uses of Chi Square test.

4. In a survey of 200 boys of which 75 were intelligent, 40 had educated fathers, while 85 of the un intelligent boys had uneducated fathers . Do these data support the hypothesis that educated fathers have intelligent boys?
Table value of chi

5. Two treatments were tried out in a control of certain type of a plant infection and with the following results:-

Treatment A : 400 plants examined and 80 found infected.

Treatment B : 400 plants examined and 70 found infected.

May it be concluded that treatment A is Superior to Treatment B ? Table value of chi square for 1 d.f = 3.841

6. In an experiment on immunization of cattle from tuberculosis, the following results were obtained:-

Treatment	Affected	Unaffected
Inoculated	31	469
No Inoculated	185	1,315

Test the effectiveness of immunization in preventing tuberculosis. The table value of Chi-square at 5% level of significance for 1 d.f is 3.84.

REFERENCE

7. **Lind, Marchal, Wathen (or Keller)**: *Basic Statistics for Business & Economics* / *Statistical Techniques in Business & Economics*.
8. **Doane, David F.**: *Essential Statistics in Business & Economics*.
9. **Spiegel, Murray R**.: *Statistics* (Schaum's Outline Series) – Great for foundational concepts.
10. **Srivastava & Rego**: *Statistics for Management*
11. **Goon, Gupta & Dasgupta**: *Fundamentals of Statistics* (For deeper theory).
12. **JK Thukral**: *Business Statistics*