Yashwantrao
Chavan
Maharashtra
Open University

Business Statistics

Author: Dr. Prashant Suresh Salve

BLOCK I: FOUNDATIONS OF STATISTICS

UNIT 1: Introduction to Business Statistics

1.1 Origin and Meaning of Statistics
1.2 Purpose of Statistics
1.3 Scope and Limitations of Statistics
1.4 Statistics: Science or Art?
1.5 Trust of Statistics
1.6 Importance of Statistics in Business Decision Making

UNIT 2: Collection and Presentation of Data

2.1 Sources of Data
2.2 Methods of Data Collection
2.3 Principles of Data Classification
2.4 Tabulation of Data
2.5 Data Presentation Techniques

BLOCK II: DESCRIPTIVE STATISTICS

UNIT 3: Frequency Distributions and Measures of Central Tendency

3.1 Frequency Distribution
3.2 Graphic Representation of Frequency Distributions
3.3 Measures of Central Tendency
    3.3.1 Arithmetic Mean
    3.3.2 Geometric Mean
    3.3.3 Harmonic Mean
    3.3.4 Mode
    3.3.5 Median
3.4 Merits and Demerits of Mean, Mode, and Median
3.5 Applications of Central Tendency in Business

UNIT 4: Measures of Variation, Skewness, and Dispersion

4.1 Measures of Variation
4.2 Skewness
4.3 Dispersion
4.4 Range, Variance, and Standard Deviation
4.5 Applications of Variation and Skewness in Business Analysis

UNIT 5: Introduction to Correlation and Regression Analysis

5.1 Definition and Significance of Correlation and Regression
5.2 Simple Linear Regression Model
5.3 Coefficients of Regression
5.4 Applications of Correlation and Regression in Business

UNIT 12: Index Numbers

12.1 Characteristics and Utility of Index Numbers
12.2 Methods of Constructing Index Numbers
12.3 Problems in Construction of Index Numbers
12.4 Limitations of Index Numbers
12.5 Consumer Price Index and Wholesale Price Index

BLOCK V: ADVANCED STATISTICAL METHODS

UNIT 13: Non-Parametric Tests

13.1 Introduction to Non-Parametric Tests
13.2 Chi-Square Test
13.3 Mann-Whitney U Test
13.4 Kruskal-Wallis Test
13.5 Wilcoxon Signed-Rank Test

UNIT 14: Multivariate Analysis

14.1 Introduction to Multivariate Analysis
14.2 Factor Analysis
14.3 Cluster Analysis
14.4 Principal Component Analysis
14.5 Discriminant Analysis

UNIT 15: Statistical Quality Control

15.1 Concept of Statistical Quality Control
15.2 Control Charts for Variables
15.3 Control Charts for Attributes
15.4 Process Capability Analysis
15.5 Six Sigma Methodology

UNIT 16: Decision Theory and Applications

16.1 Introduction to Decision Theory
16.2 Decision Making under Uncertainty
16.3 Decision Making under Risk
16.4 Decision Trees and Payoff Tables
16.5 Applications of Decision Theory in Business Strategy

Unit 1: Introduction to Business Statistics

Learning Outcomes

- Students will be able to define the meaning and purpose of statistics.
- Students will be able to evaluate the scope and limitations of statistics.
- Students will be able to analyse whether science is a science or an art.
- Students will be able to understand three dimensions of trust in statistics.
- Students will be able to remember the importance of trusting Statistics in business decision-making.

Structure

1.1 Introduction to Business Statistics

- **Origin of Statistics**

A statistician utilises historical data to estimate the likelihood that will occur rather than assuming that the dice are fair. The word "statistics" was first used centuries ago to describe the information that kings needed on their territory, crops, people, and armed forces. However, throughout time, the meaning of the word statistics has evolved significantly. Girolamo Cardano computed various dice roll probabilities in the sixteenth century.

- **Meaning of statistics**

The Latin word “status” OR the Italian word “statista” are the sources of the term “statistics.” Prof. G. F. Achenwall first used it in the eighteenth century. In the beginning, these terms were used to describe the political situation in the area. The term “statista” in Italian refers to the person who keeps track of census records or information on a state’s or country’s wealth.

According to Sir R. A. Fisher, “The science of Statistics is essentially a branch of applied mathematics and may be regarded as mathematics applied to observational data”.

Statistics’ use and significance have grown over time, and as a result, so has their nature. “Statistics” to the average person refers to numerical data stated in quantitative phrases, which may be associated with things, people, things being done, data, events, or areas of space. Because the term “statistics” may suggest various things in both singular and plural senses, there are two broad definitions for it.

1.2 Scope and Limitations of Statistics

Scope of Statistics

In recent decades, statistics has become commonly integrated into several fields, such as agriculture, biology, business, social sciences, engineering, medicine, and more. Statistical approaches are often employed to analyze and interpret experimental data. Furthermore, the extensive and diverse range of uses has resulted in the emergence of several new disciplines within the field of statistics, including Industrial Statistics, Biometrics, Biostatistics, Agricultural Statistics, and the recently formed Statistical Bioinformatics.

Statistics is a scientific discipline that involves converting data into meaningful information. The primary responsibility of statisticians is to contribute to the advancement of research and society by developing, understanding, and sharing

cutting-edge methods for gathering, presenting, evaluating, and making conclusions based on data.

Limitations of statistics

The 5 main limitations of statistics are:

- **Neglected Qualitative Aspect:** The nature of phenomena that cannot be quantified is not studied using statistical approaches. Such events are not understandable to statistical analysis. They consist of wealth, health, knowledge, etc. Qualitative data must be transformed into quantitative data.
- **It doesn't handle specific things:** According to Prof. Horace Sacrist's definition, "By statistics, we mean aggregates of facts… and placed in relation to each other," it is evident that statistics only deals with groups of facts or things; it does not identify any particular object.
- **It does not provide the full replica of the phenomenon:** When a phenomenon occurs, it is often the result of several factors, none of which can be quantified. We are unable to draw the appropriate inferences.

- **Knowledge Check-1**

  Fill in the blanks

  1. It is widely acknowledged that probability theory was founded by_____. (Kolmogorov)
  2. _____ Statistics is the process of making judgements based on data. (Inferential)
  3. "Statistics" to the average person refers to numerical data stated in _____ phrases, which may be associated with things, people, things being done, data, events, or areas of space. (Quantitative)
  4. In the _____ century, DeMoivre observed that the binomial pdf grew closer to a very smooth curve as the number of coin flips rose. (18^{th})

- **Outcome Based Activity -1**

  You manage a brand-new ice cream shop! Your chocolate chip cookie dough is popular; however, vanilla seems to be selling more slowly. Can you determine how

many scoops of each flavour to prepare tomorrow using statistics? Explain your response.

1.3 Statistics: Science or Art

Statistics is a subject of frequent debate, as is whether it is a science or an art. Not everything can be categorised as "science" or "art." Statistics is both an art and a science. A subfield of mathematics called statistics is concerned with models and procedures for data analysis.

Statistics is Science

- Because statistical methods are systematic and have many uses, statistics is a science.
- Since statistics is a science, it must be used in the same manner as any other method. Statistics is the science of gathering, organising, presenting, analysing, and interpreting numerical data to help people make better decisions.
- You must first decide what you hope to gain from statistics before deciding how to use them. For instance, the consensus is that no amount of statistical data will provide the intended outcome if the data are of insufficient quality.
- It is comparable to science since data analysis techniques frequently start with explicit models. Its final use depends on whether the model chosen is appropriate and, if so, how successfully the model is applied. This is a similar challenge to science (and engineering) in the application.

Statistics is an Art

- It is a well-known fact that art is action, while science is knowledge. From this angle, statistics may be seen as an art form. It involves using the specified techniques to gather information, conclude, and then use those conclusions to take appropriate action.
- Expertise and ability with statistical tools are necessary for the successful implementation of these qualities, which degrade art.
- Everything that is created, manufactured, or performed by hand is considered art. It is intended to be an interpretative process, however, complete with ambiguities and approximations.

Statistics, both as science and art

- A real understanding of mathematics is necessary to pursue a career in statistics, which is a field of mathematics. The basis of practically all scientific disciplines of study and investigation, mathematics is both an art and a science.
- One area of mathematics is theoretical statistics. It is actually a math problem, not an artistic or scientific one. While interpretation is an art form, maths is a science.
- Data analysis is a science and an art. Without a doubt, mathematics is a science, and the way data is visualised, for instance, may qualify as art.

1.4 Trust of Statistics

- **Three Dimensions of Trust**

The study of incomplete data is known as statistics. Statistics are mostly approximations derived from accessible and reasonably priced sources, such as administrative data initially gathered for different reasons or sample surveys.

Three aspects of trusting statistics are examined:

1) We should be able to rely on the statistical agency to provide us with accurate information about the data in a helpful manner;
2) We should be able to rely on it to produce the best estimates within reasonable timeframes and resource constraints and
3) We should be able to rely on it to decide which statistics to produce, in the sense of having planning systems that are adaptable to society's changing data needs.

- **Importance of trusting Statistics**

Statistics are essential for guiding decisions and building our thinking in the data-driven world. Here's why it's crucial to believe statistics:

- Making sophisticated decisions: Statistical analysis is essential for making many important decisions, ranging from public policy to personal health. Decisions based on unreliable statistics may be erroneous and have unfavourable outcomes. Imagine if a physician based their care on inaccurate information on the effectiveness of a drug.
- Public confidence and welfare: Societal conditions are assessed by official statistics of crime or unemployment with which governments and individuals work. When the level of public trust weakens in these personalities, people will cease to engage

in activities within the public domain or doubt the impact that the government policies hold.

1.5 Importance of statistics in business decision-making

Statistics is not anymore a specialisation area in today's fast-changing corporate world, and globalisation is totally based on data support; it is an essential tool of decision making. By using statistics, firms are able to move away from pure guesswork and decisions that may not be very effective. They guide how to analyse data, make trends, and gain other significant knowledge that will be helpful in strategic management. They explore the possibility of examining the rationale behind the subject of statistics and its role in the effectiveness of the companies.

1) Data-driven decision-making: Qualitative sales data, specific customer details, the flow of traffic on corporations' websites, and other business information are abundant in the corporate world. Statistics offers tools and techniques for this purpose, that is, arranging, examining, and analysing this data to get useful solutions or conclusions.
2) Risk assessment and justification: Many business choices involve essential possibilities. Businesses may estimate risks and create strategies to reduce them using statistics.
3) Determining Customer Preferences and Industry Trends: Gaining insight into consumer behaviour and industry trends is essential for company expansion. Businesses may identify developing trends, client preferences, and possible new markets by using statistical approaches such as customer segmentation analysis and market research surveys.

- **Knowledge Check 2**

  State True or False

  1. Statistics are essential for guiding decisions and building our thinking in the data-driven world of today. (True)
  2. Statistics is neither a science nor an art. (False)
  3. Today, statistics are employed in practically every scientific field, even though a century ago, most natural scientists were unlikely to acknowledge statistics as a distinct discipline. (True)
  4. There are six dimensions of trust in statistics. (False)

- **Outcome-Based Activity -2**

  As a social media manager, you are curious about the effectiveness of your most recent video advertisement. You have a view, like, and comment data. Describe how statistics is an art as well as a science that can help you evaluate this data and determine the effectiveness of the advertising strategy.

1.5 Summary

- A statistician utilises historical data to estimate the likelihood that a face will fall face up rather than assuming that the dice are fair.
- The Latin term status, from which the English word statistics is derived, signifies "political state" or "government."
- The term "statista" in Italian refers to the person who keeps track of census records or information on a state's or country's wealth.
- Because the term "statistics" may signify various things in both singular and plural senses, there are two broad definitions for it.
- The goal of descriptive analysis is to summarise unprocessed data on the business, its clients, its goods, etc.
- Probability aids in calculating the possibility of upcoming events.

1.6 Keywords

- **Inferential Statistics:** It is the process of making judgements based on data.
- **Descriptive Statistics:** It is concerned with summarising and characterising data.
- **Statista:** It refers to the person who keeps track of census records or information on a state's or country's wealth.
- **Statistics:** It refers to numerical data stated in quantitative phrases, which may be associated with things, people, things being done, data, events, or areas of space.

1.7 Self-Assessment Questions

1. What is the origin of Statistics?
2. What are the scope and limitations of statistics?
3. What do you infer? Is statistics a science or an art?
4. What do you mean by Trust in statistics, and what is its importance?

1.8 References/ Reference Reading

- Gupta, S. C., and V. K. Kapoor. *Fundamentals of Mathematical Statistics*. Sultan Chand & Sons, 2020.
- Levin, Richard I., and David S. Rubin. *Statistics for Management*. Pearson Education, 2021.
- Goon, A. M., M. K. Gupta, and B. Dasgupta. *Fundamentals of Statistics*. Vol. 1, World Press Private Ltd, 2019.
- Beri, G. C. *Business Statistics*. McGraw Hill Education, 2018.
- Spiegel, Murray R., Larry J. Stephens, and Narinder Kumar. *Schaum's Outline of Business Statistics*. 4th ed., McGraw Hill Education, 2020.

Unit 2: Collection and Presentation of Data

Learning Outcomes

- Students will be able to define the sources of data collection.
- Students will be able to evaluate the methods of Collecting Primary and secondary data.
- Students will be able to analyse the principles of data collection.
- Students will be able to understand the principles of Tabulation.
- Students will be able to remember data presentation techniques in statistics.

Structure

2.1 Sources of Data

- **Meaning of Data**

Data is a set of measurements and facts that may be used as a tool to provide information to help a person or group of individuals arrive at a solid decision. It aids in the analyst's comprehension, analysis, and interpretation of many socioeconomic issues, such as inflation, unemployment, and poverty.

Sources of Collection of Data

Primary source

It is a compilation of information from the source. It gives the researcher access to first-hand, quantitative, unprocessed data about the statistical analysis. To put it briefly, direct access to the subject is provided via the primary sources of data.

Secondary Source

It is a compilation of information gathered from many organisations that originally obtained the information from sources. It does not give the researcher access to first-hand, quantitative, unprocessed data on the study.

While primary sources lend greater credibility to the data they collect because they offer supporting evidence, a thorough research project will need to collect data from both primary and secondary sources.

- **Primary and Secondary Data**

Primary Data

Primary data are those that are initially gathered from scratch by the investigator from primary sources. This information was collected straight from the source. It is data that is current and always tailored to the requirements of the researcher. The raw version of the main data is accessible. The investigator must commit a significant amount of time and money to gather primary data.

Secondary Data

Secondary data is information that already exists but was previously gathered for a different reason by another party. Since such material has previously been the subject of investigation, it does not contain any real-time data. However, gathering secondary data comes at a lower cost. The data is available in a refined version since it has previously been collected. Compared to main data, secondary data has comparatively lower accuracy and dependability.

2.2 Methods of Data Collection

- **Meaning of Data Collection**

The practice of gathering information from pertinent sources to address a particular statistical question is known as data collection. The most important stage in any statistical inquiry is data collection.

- **Methods of Collecting Primary Data**

1) **Direct Personal Investigation:** This approach, as the name implies, involves gathering information directly from the source of origin. In other words, the investigator speaks with the subject directly to obtain information.
2) **Indirect Oral Investigation:** This technique of gathering primary data involves the investigator speaking orally with someone who possesses the desired information rather than going directly to the person from whom the information is needed.
3) **Information from Local Sources or Correspondents:** In this approach, the investigator assigns local individuals or correspondents to different locations to gather data, which they then provide to the investigator.

- **Methods of Collecting Secondary Data**

1) Printed Sources

a) Publications of the Government: As part of its regular business, the Indian government publishes a variety of publications that include various types of information or data released by the Ministries, Central, and State governments.
b) Semi-Government Publications: A variety of semi-government organisations distribute information on births, deaths, health, and education.
c) Trade Association Publications: A number of large trade associations gather and spread information about various elements of commercial activity from their statistics and research sections.

2) Unpublished Sources

Unpublished sources are another place to get secondary data. These sources are gathered by several governmental agencies and other organisations. In general, these groups collect information for their purposes and do not transfer it.

- **Knowledge Check -1**

  Fill in the blanks

  1) _____ is a set of measurements and facts that may be used as a tool to provide information to help a person or group of individuals arrive at a solid decision. (Data)
  2) _____ is a compilation of information from the source. (Primary Source)
  3) A _____ is an individual from whom the statistical data needed for the investigation is gathered. (Respondent)
  4) A _____ primary goal is to gather information about several attributes, including pricing, friendliness, quality, and usefulness. (Survey)

- **Outcome Based Activity -1**

  You should find out if your class's pupils would rather work on solo or group assignments. How would you gather information to respond to this query? Summarise the data-gathering technique you have selected.

2.3 Principles of Data Collection

The thorough gathering of data is the foundation of every statistical analysis and ensures the reliability and use of the information that is produced. Here are some essential guidelines to remember:

1) Organising is Essential:

This perspective makes a strong point that one needs to have a clear plan for how the collection will be done before using this technique. Preparation works as follows:

A) Concentrate your efforts: In simple words, if you provide another well-defined problem to search for the data, you will get the significant data points relevant to your study.

B) Data collecting technique guidance: This shows that no extended approach can be used to collect data because the strategy used will depend on the type of data being collected. If you have a chief concern on thoughts, surveys might be the best option for you.

C) Boosts productivity: It allows us to avoid spending time and money collecting data, which is often very unnecessary information.

2) Maintain Simplicity:

This guideline only allows for your data-collecting mechanisms; usability is selected. This is how simplifying things enhances the gathering of data:

A) Asking specific short questions: Avoid such questions that you think may be answered using extremely technical language. Ensure your words are clear for people to comprehend, and give an appropriate response to the spoken words.
B) Ensure that you adopt an effective method of data collection depending on the type of information you require: Quantitative research strategies, such as self-completed questionnaires, include closed questions and may involve the use of multiple-choice and are better suited for collecting data in a short space of time, as compared to qualitative research methodologies.
C) Reduce the workload for participants: In this context, try to achieve a suggested, efficient procedure for data gathering that will impose minimum time or enervation on participants. It opens up an opportunity to receive correct and detailed responses to a given question.

2.4 Tabulation of Data

- **Meaning of Tabulation of Data**

Tabulation is the process of organising the recorded facts and data into a tabular format. A table is an arrangement of rows and columns that symmetrically shows arithmetical facts. While columns are constructed and displayed perpendicularly, rows are prepared horizontally. Tabulation is a technique that allows an examiner to quickly and methodically present the needed information by summarising grouped or ungrouped data in a tabular style that is easy to grasp. A statistical table enables the researcher to show a vast amount of data or information methodically and thoroughly. It makes association easier and frequently makes patterns in data visible that might otherwise go undetected.

- **Types of Tabulation**

Three different types of tabulation exist.

- Basic Tabulation
- The Pairwise Comparison and
- Intricate Tablework (merging Cross-tabulation)

1) Basic Tabulation: The data are tallied using a single, familiar format.

Pilot research conducted in a classroom on July 7, 2017, clearly demonstrated the frequency with which all of the chosen pupils had pens from various brands, including Parker, A.T. Cross, Aurora, Mont Blanc, Cello, Reynolds, Camlin, Sheaffer, Papermate, Shanghai Hero, and so on.

One – Way Table	
Brands of Pens	**No. of Students (Class wise/Division)**
Monte Blanc	40
Cello	50
Reynolds	65
Camlin	38
Sheaffer	40
Shanghai Hero	43
Total	**276**

2) **Double tabulation involves tabulating two distinct or exclusive sets of data. For example, each class has a uniform number** of boys and girls who have laptops from different brands, such as Apple, IBM, HP, Lenovo, Sony, Acer, and Dell.

Laptop Brands	No. of Students (Gender wise)		Total
	Male	Female	
Apple	06	02	08
IBM	07	05	12
Lenovo	10	11	21
HP	12	09	21
Acer	20	15	35
Sony	12	13	25
Total	**67**	**55**	**122**

3) **Complex Tabulation:** More than two properties are included in the complex tabulation of figures.

For example, on July 16, 2017, pilot research was conducted to find out how many students, both males and girls, in the 21- to 23-year-old age range had laptops of a certain brand. The primary goal of this is to cross-tabulate the data, which may include information on boys and girls who may own specialised laptops because they belong to a particular age group.

Laptop Brands	No. of Students [Age group wise (A)]											
	Male				Female				Total			
	21(A)	22(A)	23(A)	Total	21(A)	22(A)	23(A)	Total	21(A)	22(A)	23(A)	Tota
Apple	1	2	3	06	2	3	4	09	5	6	1	12
IBM	2	4	6	12	4	6	2	12	2	2	2	06
Lenovo	3	5	7	15	5	7	3	15	2	1	3	06
HP	3	2	1	06	2	1	3	06	5	7	3	15
Acer	2	2	2	06	2	2	2	06	4	6	2	12
Sony	4	5	6	15	5	6	1	12	2	3	4	09
Total	**15**	**20**	**25**	**60**	**20**	**25**	**15**	**60**	**20**	**25**	**15**	**60**

- **Merits and Limitations of Tabulation**

Merits

1) It makes understanding complicated information simple.
2) Comparing similar information will be made easier, and it's easier to compute other statistical measures like averages, dispersion, correlation, etc.

Limitations

1) Only numerical facts and qualitative expressions are found in tables.
2) When conclusions are reached that are too complex for the average layperson to grasp, tables might help.

2.5 Data Presentation Techniques

Just as important as the analysis itself is the capacity to communicate statistical findings effectively. Below is a summary of several important statistical data display techniques:

i) Tables: The table is a fundamental tool for arranging structured data into rows and columns. It is very helpful for showing frequency distributions, group comparisons, and numerical data.

ii) Charts and graphs: Data visualisation is a very effective tool for communicating trends, patterns, and correlations between variables.

- **Knowledge Check 2**

  State True or False

  1) Tables must be clear, concise, and well-labelled. (True)
  2) Tabulation is the process of organising the recorded facts and data into a horizontal format. (False)

3) More than two attributes are included in the complex tabulation of figures. (True)
4) Double tabulation involves tabulating four distinct or exclusive sets of data. (False)

- **Outcome-Based Activity 2**

  Your business expects to publish a brand-new social networking app. Which two criteria would you apply when categorising the information that users will submit (such as names, images, and messages)? Give a brief justification for your decisions.

2.6 Summary

- Data is a set of measurements and facts that may be used as a tool to provide information to help a person or group of individuals arrive at a solid decision.
- It gives the researcher access to first-hand, quantitative, unprocessed data about the statistical analysis.
- Secondary data is information that already exists but was previously gathered for a different reason by another party.
- Since the primary data was gathered straight from the source of origin, it is unique.
- However, because secondary data is acquired from both public and unpublished sources, it is less expensive to collect.
- Quantitative data sources employ numbers, but qualitative data sources do not.

2.7 Keywords

- **Data:** It is a set of measurements and facts that may be used as a tool to provide information to help a person or group of individuals arrive at a solid decision.
- **Primary Sources:** It is a compilation of information from the source.
- **Secondary Data:** Information that already exists but was previously gathered for a different reason by another party.
- **Respondent:** A respondent is an individual from whom the statistical data needed for the investigation is gathered.
- **Direct Personal Investigation**: This approach involves gathering information directly from the source of origin, as the name implies.

- **Tabulation:** It is the process of organising the recorded facts and data into a tabular format.

2.8 Self-Assessment Questions

1. What are the different sources of data collection?
2. What are the different methods of Collecting Primary and Secondary Data?
3. What are the different principles of data classification?
4. What are the different types of Tabulation?
5. What are the merits and limitations of Tabulation?

2.9 References/ Reference Reading

- Gupta, S. C. "Fundamentals of Statistics." Himalaya Publishing House, 2022.
- Sharma, J. K. "Business Statistics." Vikas Publishing House, 2021.
- Arora, P. N., and S. Arora. "Statistics for Management." S. Chand & Company Pvt. Ltd., 2020.
- Kothari, C. R., and Gaurav Garg. "Research Methodology: Methods and Techniques." New Age International Publishers, 2019.
- Miller, Jane E., and N. S. Raju. "Data Presentation Techniques for Business." Pearson Education India, 2022.

Unit 3: Frequency Distributions and Measures of Central Tendency

Learning Outcomes:

- Students will be able to define the meaning of frequency distribution.

- Students will be able to understand the merits of mean, mode and Median.
- Students will be able to remember the demerits of mean, mode and Median.
- Students will be able to analyse a graphical representation of frequency distribution.
- Students will be able to apply the different measures of Central tendency.

Structure

3.1 Frequency Distribution

- **Meaning of frequency distribution**

A frequency distribution is used to arrange the collected data into tables. Grades, town temperatures, volleyball focus scores, etc., can all be examples of the information.

Let's look at an example to help clarify this. The results of ten pupils on the G.K. quiz that Mr. Chris provided are as follows: 15, 17, 20, 15, 20, 17, 17, 14, 14, 20. Using a frequency distribution to depict this data, we can determine how many students received identical grades.

Quiz Marks	**No of students**
15	2
17	3
20	3
14	2

- **Types of frequency distribution**

Under statistics, there are four different forms of frequency distribution, which are described below:

a) **Ungrouped frequency distribution:** Instead of displaying data value groupings, it presents the frequency of each item in each data value.

b) **Grouped frequency distribution:** This class divides and arranges the data into units known as class intervals. A frequency distribution table shows the frequency of the data for each class interval. The distribution of frequencies in class intervals is displayed in the grouped frequency table.

c) **The relative frequency distribution:** This indicates the percentage of all observations that fall into each group.

d) **Cumulative frequency distribution:** The total of all frequencies in a frequency distribution that is below the initial frequency is known as the cumulative frequency distribution.

3.2 Graphic Representation of Frequency Distribution

- **Meaning of Graphic Representation of Frequency Distribution**

Vast data sets might be intimidating. Frequency distributions, which display the frequency at which each value or range of values occurs, aid in organising this data.

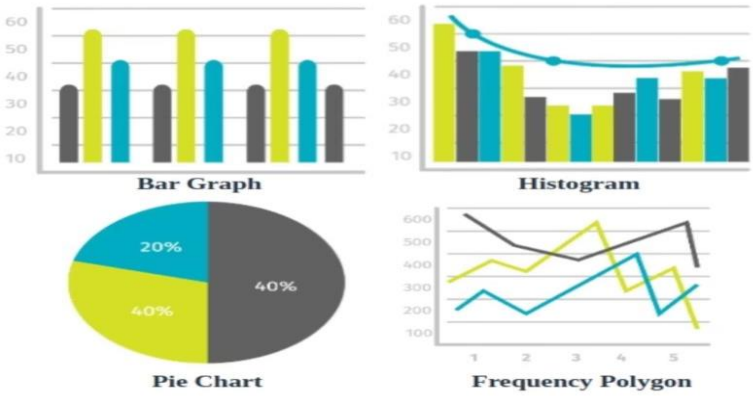
For this, several popular graph types are used, each with advantages of its own:

A) **Histograms:** These are bar graphs in which the bars show how frequently numerical data falls into particular ranges or classes. Since values in continuous data might fall anywhere on a spectrum, histograms are perfect for this type of data.

B) **Bar Charts:** Bar charts use bars to indicate frequencies, much like histograms do. They are usually applied to categorical data, which is not a continuous range but rather distinct categories of data points.

- **Frequency distribution graphs**

Several ways are available to depict the frequency distribution, including pie charts, bar graphs, histograms, and frequency polygons.



Graph Type	Description	Use Cases
Histogram	It uses equal-width bars to depict the frequency of each continuous data interval.	Examination of the distribution of data continuously.
Bar Graph	It uses equal-width bars to show the frequency of each interval; discontinuous data can also be represented.	I was contrasting different sorts of data.
Frequency Polygon	Connects class frequency midpoints with lines; it resembles a histogram without the bars.	Comparing various datasets

Pie Chart	Data is shown as slices of a circle on a circular graph, with each slice's proportionate size to the entire dataset indicated.	They show the relative sizes of the data portion.

- **Knowledge Check 1**

  Fill in the Blanks

  1. A _____ is utilised to arrange the gathered information into tables. (Frequency Distribution)
  2. _____ frequency distribution divides and arranges the data into units known as class intervals. (Grouped)
  3. The width of the _____ is determined by the class interval of the variable on the horizontal axis, and the corresponding frequency on the vertical axis determines the heights of the rectangles. (Histogram)
  4. Each frequency must be marked against the relevant class on the height of its corresponding ordinate in order to plot a _____. (Frequency Polygon)

- **Outcome-Based Activity 1**

  Let's say you gathered information about how many dogs each student in your class owned. To illustrate the frequency distribution of this data, make a histogram. Make sure your graph has the proper title and that the axes are labelled.

3.3 Measures of Central Tendency

- **Meaning of Central Tendency**

The mathematical qualities utilised to demonstrate the middle or central value of a substantial mixture of mathematical information are known as central tendencies in measurements. In measurements, these gained mathematical qualities are suggested to be focal or normal qualities. Since it outlines the nature or characteristics of the total

informational index, which would be incredibly difficult to see, such a number is very critical.

3.3.1 Arithmetic Mean

Meaning

The arithmetic mean, frequently known as the average, is an essential idea communicated as the sum of a bunch of numbers separated by the numbers in the set. Include every number in a given information assortment, then divide the total number of values there to view as the math mean.

Formula

The following is the formula for the mean, also referred to as the average:

The arithmetic mean is calculated as the total number of numbers in the data set divided by the sum of all the numbers. The sample arithmetic mean, represented by the symbol $\dot{x}$ (pronounced "x-bar"), has the following formula:

$$x^{-} = \sum ni=1\ xi$$

In this formula

1) The sample average, or sample arithmetic mean, is denoted by $\dot{x}$.
2) The data set's numbers are represented by xi, where I is the total number of sample observations and ranges from 1 to n.
3) Summation is indicated by the symbol ∑, which means that you should add up each integer. Xi ranging from 1 to n.
4) The total number of sample observations in the data collection is denoted by n.

3.3.2 Geometric mean

Meaning

The average value, or mean, that addresses the middle value of a bunch of numbers in math is known as the geometric mean (GM), which is determined by duplicating the values of the set. Generally, we increase the numbers by n, where n is all the complete number of information values, and afterwards, we acquire the nth foundation of the

duplicated numbers. For example, the mathematical mean of a given set of two numbers, say 3 and 1, is $\sqrt{(3\times1)} = \sqrt{3} = 1.732$.

Formula

The following is the formula to find the geometric mean:

The nth root of the product of the values is the Geometric Mean (G.M.) of a series of n observations.

Assuming that the observations are x1, x2,…, xn, the G.M. is defined as follows:

$$G.M = \sqrt[n]{x_1 \times x_2 \times \ldots x_n}$$

or

$$G.M = (x_1 \times x_2 \times \ldots x_n)^{\frac{1}{n}}$$

3.3.3 Harmonic Mean

Meaning

In statistics, the reciprocal of the arithmetic mean of the variables is known as the harmonic mean. It is thoroughly defined and grounded in all observations. It is used for times and average rates.

Formula

Formula:

If $x_1, x_2, \ldots, x_n$ are the n individual items, the harmonic mean is given by

Harmonic Mean

$$= \frac{n}{\frac{1}{x_1}+\frac{1}{x_2}+\frac{1}{x_3}+\ldots\ldots\frac{1}{x_n}}$$

3.3.4 Mode

Meaning

The worth in a given arrangement of values with the most important measurement is known as a mode. The number appears the most often. For example, since value 5 shows up two times in the given set of information (2, 4, 5, 6, 7), it is the method of the informational index. The show window, get-together, and investigation of information and data for a particular objective are the main points of measurement. We utilise

realistic depictions, tables, diagrams, pie outlines, and reference charts for measurements. To determine helpful data, the information should be appropriately coordinated and afterwards further dissected.

Formula

$$Mode = l + \left(\frac{f_1 - f_0}{2f_1 - f_0 - f_2}\right) \times h$$

Where,

l = lower limit of the modal class

h = size of the class interval

f_1 = frequency of the modal class

f_0 = frequency of the class preceding the modal class

f_2 = frequency of the class succeeding the modal class

3.3.5 Median

Meaning

The middle value or number in a bunch of information is known as the median. The number middle through the set is likewise the median. The information ought to be arranged from least to most countable or from most important to least value to get the middle. A number that separates the upper portion of a population, information test, or likelihood conveyance from the lower half is known as the middle. There are contrasts between the different conveyance types in the middle.

Formulas

A) Odd no of observations

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

where n is the number of observations

B) Even no of observations

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

where n is the number of observations

3.4 Merits and Demerits of Mean, Mode and Median

- **Merits of Mean, Mode and Median**

Mean

1**) Simple and easy calculation:** To calculate the average, add up all the values in the data set, then divide the total number of values by the number of values. This method is used by beginners and is very popular because of its simplicity.

2) **It examines each data point:** The mean measures each value in the data set, unlike the mode. Compared to simply finding the most frequent value, this gives a more complete picture of the general trend.

Mode

1) **Simple and easy to calculate:** The concept of a mode is really simple. It's easy to capture and calculate by hand or simple software—all you have to do is set the most likely value.

2) **Not affected by extremes:** Unlike the mean, the mode is not affected by extreme values in the data set. This makes it a smarter decision if the data distribution is skewed.

3) **Not only numerical data:** the mode can also be applied to categorical data.

Median

1) **Outlier-resistant:** The median is unaffected by extreme values in the data set, in contrast to the mean.
2) **Simple to comprehend and compute:** To obtain the median, arrange the data and identify the middle number.

- **Demerits of Mean, Mode and Median**

Mean

1) Sensitive to outliers: The extreme values in the data set have a significant impact on the mean. A single extreme can greatly change the average and misrepresent the typical data point.
2) Not relevant to data that is categorical: Only numerical data sets can use the mean.
3) Disregard the distribution's foundation: The mean does not show the data's form or distribution.

Mode

1) Uncertain in some situations: When all values occur equally frequently or when there are many values with the same greatest frequency (bimodal or multimodal data), the mode may be unclear.
2) Not comprehensive: The mode does not take into account every data point in the set, unlike the mean or median. Only the most frequent value(s) is/are highlighted.

Median

1) Ignores certain data: The median does not use all of the information in the data set; it just takes into account the middle value or values.
2) Not as representative: When dealing with skewed data that has extreme values, the median may not correctly represent the "typical" data point in comparison to the mean.

3.5 Applications of Central tendency in business

Measures that condense the "average" or "middle" of a data set are referred to as central tendency, a fundamental idea in statistics. Understanding mean, median and mode is essential for making well-informed judgements in various sectors within the corporate world.

1) Marketing and Sales: By defining the "typical" consumer, central tendency measurements direct marketing initiatives.
2) Measures of central tendency are essential for maximising inventory levels in operations and inventory management.

- **Knowledge Check 2**

  State True or False

  1. The median may not correctly represent the "typical" data point in comparison to the mean when dealing with skewed data that has extreme values. (True)

2. In statistical terms, the mean divides your data set in half, serving as a centre point. (False)
3. The value in a given set of values with the highest frequency is called a mode. (True)
4. The mean shows the form or distribution of the data. (False)

- **Outcome-Based Activity 2**

  Assume you are a bakery owner who wants to know what kinds of cupcakes your clients enjoy. You compile information on how many cupcakes each client bought throughout the previous week. Determine the data's mean, median, and mode. Give a brief explanation of which central tendency measure and why best captures the average number of cupcakes purchased.

3.6 Summary

- A frequency distribution is utilised to arrange the gathered information into tables.
- In essence, these graphical depictions are the frequency distribution table summarised visually.
- The display, gathering, and analysis of data and information for a specific goal is the focus of statistics.
- The mode makes it simple to determine which item, size, preference, or result in your data is the most popular by providing you with a clear image of the most often occurring value.

3.7 Keywords

- **Frequency distribution** – A frequency distribution is utilised to arrange the gathered information into tables.
- **Histogram** – These are bar graphs in which the bars show how frequently numerical data falls into particular ranges or classes
- **Central Tendency**—In statistics, the numerical values used to indicate the mid-value or middle value of a sizable collection of numerical data are known as central tendencies.

- **Arithmetic Mean-** The arithmetic mean, often known as the average, is a basic concept that is expressed as the sum of a set of numbers divided by the total number of numbers in the set
- **Mode**- The value in a given set of values with the highest frequency is called a mode.
- **Median-** The middle value or number in a set of data is called the median.

3.8 Self-Assessment Questions

1. What do you mean by frequency distribution?
2. What are the different types of Graphical representations of frequency distribution?
3. What do you mean by measures of Central tendency?
4. What are the different measures of Central tendency?
5. What are the merits and limitations of mean, mode and Median?

3.9 References/ Reference Reading

- Gupta, S. P., and M. P. Gupta. Business Statistics. Sultan Chand & Sons, 2022..
- Levin, Richard I., and David S. Rubin. Statistics for Management. Pearson Education India, 2017.
- Sharma, J. K. Business Statistics: Problems and Solutions. Pearson Education India, 2018.
- Vohra, N. D. Business Statistics. McGraw Hill Education (India), 2017.
- Spiegel, Murray R., Larry J. Stephens, and Narinder Kumar. Schaum's Outline of Theory and Problems of Statistics. McGraw Hill Education (India), 2017.

Unit 4: Measures of Variation, Skewness, and Dispersion

Learning outcomes

- Students will be able to evaluate the significance of measuring variation.
- Students will be able to analyse different measures of variation in statistics.
- Students will be able to apply the formula of relative Skewness.
- Students will be able to understand the types of dispersion in statistics.
- Students will be able to remember the difference between measures of dispersion and central tendency.

Structure

4.9 References/ Reference Reading

4.1 Measures of variation

- **Meaning of measure of variation**

The spreading of the singular qualities around the middle worth is depicted by a proportion of variation, otherwise called dispersion. We should check out the accompanying information to exhibit the possibility of variation:

Firm A Daily Sales (Rs.)	Firm B Daily Sales (Rs.)	Firm C Daily Sales (Rs.)
5000	5050	4900
5000	5025	3100
5000	4950	2200
5000	4835	1800
5000	5140	13000
$\bar{X}_A = 5000$	$\bar{X}_B = 5000$	$\bar{X}_C = 5000$

- **Quartile Deviation**

The difference between the third and first quartiles is averaged to get the quartile deviation, sometimes referred to as the semi-interquartile range. This can also be expressed in symbols as:

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

where Q_1 = first quartile, and Q_3 = third quartile.

The following illustration will clarify the procedure involved. For the data given below, compute the quartile deviation.

Monthly wages (Rs.)	No of workers	Monthly wages (Rs.)	No of workers
Below 850	12	1000-1050	62
850-900	16	1050-1100	75
900-950	39	1100-1150	30
950-1000	56	1150& above	10

To compute quartile deviation, we need the values of the first quartile and the Third quartile, which can be obtained from the following table:

Monthly wages (Rs.)	No of workers	C.F.
Below 850	12	12

850-900	16	28
900-950	39	67
950-1000	56	123
1000-1050	62	185
1050-1100	75	260
1100-1150	30	290
1150& above	10	300

Q_1 = Size of $\frac{N}{4}$th observation $= \frac{300}{4} = 75$th observation which lies in the class $950 - 1000$

$$Q_1 = L + \frac{N/4 - pcf}{f} \times i = 950 + \frac{75 - 67}{56} \times 50$$

$$= 950 + \frac{50}{7} = 950 + 7.14 = 957.14$$

Q_3 = Size of $\frac{3N}{4}$th observation $= \frac{3\times300}{4} = 225^{th}$ observation which lies in the class $1050 - 1100$

$$Q_3 = L + \frac{3N/4 - pcf}{f} \times i = 1050 + \frac{225 - 185}{75} \times 50$$

$$= 1050 + \frac{2000}{75} = 1050 + 26.67 = 1076.67$$

$$Q.D. = \frac{1076.67 - 957.14}{2} = \frac{119.53}{2} = 59.765$$

The relative measure corresponding to quartile deviation, called the coefficient of quartile deviation, is calculated as given below:

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

- **Average Deviation**

Because it takes into account every observation in the provided set of data, the average (or mean) deviation improves upon the previous two measures. The average of deviations from the mean or the median is used to calculate this metric. Regardless of the sign, all deviations are seen as positive. This can be expressed symbolically by:

$$A.D. = \frac{\sum|X - \bar{X}|}{N} \text{ or } \frac{\sum | X - \text{Median} |}{N}$$

In theory, there is a benefit of taking the deviations from the median since the total of the absolute deviations—that is, the deviations from the median without accounting for ± signs—is the least. The following formula is to be used to group data:

$$A.D. = \frac{\sum|X - \bar{X}|}{N}$$

- **Coefficient of Variation**

The coefficient of variation contracted C.V. is a usually utilised relative proportion of change. This measurement is just the standard deviation to the mean proportion given as a rate.

Consider the additional information that connects with the mean everyday deals and standard deviation for four locales.

Region	**Mean daily sales(Rs. 1000)**	**Standard deviations (Rs.1000)**
1	86	10.45
2	45	5.86
3	72	72
4	61	11.32

We will compute the coefficients of variation to ascertain which region has the most constant daily sales. You might have noticed that each region's mean daily sales are not the same.

$$C.V._1 = \frac{10.45}{86} \times 100 = 12.15; C.V._2 = \frac{5.86}{45} \times 100 = 13.02$$

$$C.V._3 = \frac{9.54}{72} \times 100 = 13.25; C.V._4 = \frac{11.32}{61} \times 100 = 18.56$$

As the coefficient of variation is the minimum for Region 1, the most consistent region is Region 1.

4.2 Skewness

- **Meaning of Skewness**

Not every feature of a particular collection of data may be discovered by the metrics of central tendency and variance. The guidelines are:

A) When low and high numbers balance each other out, and there are no extreme values in one direction, the data are said to be symmetrical. Here, mean equals median equals mode.

B) The skewness is negative if the longer tail is pointing towards the left or lower value. When some really low numbers lower the mean, the result is negative skewness, where mean < median < mode.

- **Relative Skewness**

The Karl Pearson-provided coefficient of skewness may be defined as follows in order to compare the skewness in two or more distributions:

$$SK. = \frac{\text{Mean - Mode}}{\text{S. D.}}$$

If the mode cannot be determined, then using the approximate relationship, Mode = 3 Median - 2 Mean, the above formula reduces to

$$SK. = \frac{3\ (\text{Mean - Median})}{\text{S.D.}}$$

The conveyance is balanced on the off chance that the coefficient's worth is zero; emphatically slanted dispersions have positive coefficient values; adversely slanted circulations have negative.

$$SK. = \frac{Q_3 + Q_1 - 2\text{Median}}{Q_3 - Q_1}$$

Coefficient values. In common-sense application, this coefficient's worth ordinarily falls between ± 1. The coefficient of skewness recipe given by Bowley is more reasonable when managing open-ended circulations that contain extreme qualities in the information or situating measurements like the middle and quartiles.

Again, this distribution is symmetrical if the value of this coefficient is 0.

It is a positively skewed distribution for positive values and a negatively skewed distribution for negative values. The application of Bowley's method would be clear by considering the following data:

Sales(Rs. lakhs)	**No of companies**	**C.F.**
Below 50	8	8
50-60	12	20
60-70	20	40

70-80	25	65
80& above	15	80

Q_1 = size of $\frac{N}{4}$th observation $= \frac{80}{4}$ = 20th observation which lies in the class 50-60

$$Q_1 = L + \frac{N/4 - pcf}{f} \times i = 50 + \frac{20 - 8}{12} \times 10 = 60$$

Q_2 = Median = size of $\frac{N}{2}$th observation $= \frac{80}{2}$ = 40th observation which lies in the class 60-70

$$Q_2 = \text{Med.} = L + \frac{N/2 - pcf}{f} \times i = 60 + \frac{40 - 20}{20} \times 10 = 70$$

Q_3 = Size of $\frac{3N}{4}$th observation $= \frac{3\times80}{4} = 60$ th observation which lies in the class 70-80

$$Q_3 = L + \frac{3N/4 - pcf}{f} \times i = 70 + \frac{60 - 40}{25} \times 10 = 78$$

$$\text{Coefficient of } SK = \frac{Q_3 + Q_1 - 2\,\text{Median}}{Q_3 - Q_1}$$
$$= \frac{78 + 60 - 2 \times 70}{78 - 60} = -0.11$$

The distribution is somewhat skewed to the left, as shown by this coefficient of skewness value, and as a result, sales are concentrated more at the upper values of the distribution than at the lower values.

- **Knowledge Check -1**

  Fill in the blanks

  1) The spread or scattering of the individual values around the centre value is described by a measure of _____, also known as dispersion. (Variation)
  2) The difference between a collection of data's greatest (numerically largest) and lowest (numerically smallest) values is known as the _____. (Range)
  3) The _____ deviation Is a useful indicator of variability, but its applications are restricted. (Average)
  4) The absence of symmetry in distribution is referred to as _____. (Skewness)

- **Outcome-Based Activity 1**

  After gathering test results for Class A and Class B, you were given the following information:

  78, 82, 85, 90, 92, 95, 98 are in class A.

  65, 70, 75, 80, 80, 85, 90 in class B

  Compute and contrast the standard deviation and range for Class A and Class B. Give a brief explanation of the class with the wider score distribution.

4.3 Dispersion

- **Meaning of dispersion**

In Statistics, dispersion is a term used to show how scattered or conveyed the information is around a normal value. A range of information's distribution discovers its consistency or fluctuation. Scattering might be estimated utilising various measurements, including reach, fluctuation, and standard deviation.

- **Absolute measure of dispersion**

Absolute Measure of Dispersion refers to the dispersion metrics that are measured and expressed in the actual data units. For instance, kilogrammes, dollars, metres, etc. Several absolute dispersion measurements include:

i) Range: The difference between the distribution's greatest and lowest value determines its definition.

ii) Variance: The average square deviation of the provided data set's mean is what defines it.

iii) Quartile Deviation: In a particular data collection, it is defined as half of the difference between the first and third quartiles.

iv) Interquartile Range: The term refers to the variation between the upper (Q3) and lower (Q1) quartiles. This is its formula: Q3 – Q1.

- **Relative measure of dispersion**

To have a better understanding of the data scattering, we measure the two values with differing units using relative measures of dispersion. The following are a few examples of relative dispersion measures:

i) The ratio of the difference between the highest and lowest value in a data set to the total of the highest and lowest values is known as the coefficient of range.

ii) The ratio of the standard deviation to the data set mean is known as the coefficient of variation. To indicate the coefficient of variation, we utilise percentages.

iii) The coefficient of mean deviation is the product of the mean deviation and the value of the data set's centre point.

- **Coefficient of dispersion**

When two series with significantly different averages are compared, coefficients of dispersion are computed.

Relative Measures of Dispersion	Related Formulas
Coefficient of Range	$(H - S)/(H + S)$
Coefficient of Variation	$(SD/Mean)\times 100$
Coefficient of Mean Deviation	(Mean Deviation)/μ where, μ is the central point for which the mean is calculated
Coefficient of Quartile Deviation	$(Q_3 - Q_1)/(Q_3 + Q_1)$

4.4 Range, Variance, and Standard Deviation

Range

The difference between a collection of data's greatest (numerically largest) and lowest (numerically smallest) values is known as the range. This might be represented symbolically as follows: R = H – L, where R stands for range, H for highest value, and L for lowest value.

Consider the three companies' daily sales data from before as an example.

R = H – L = 5000 – 5000 = 0 for company A

R = H – L = 5140 – 4835 = 305 for company B

R = H – L = 13000 – 1800 = 11200 for company C

Range is a widely used notion in statistical quality control. Range analysis is useful when examining price fluctuations in commodities such as shares, debentures, and

other assets that are highly susceptible to fluctuations in value over time. The range serves as a useful weather forecasting indication for meteorological authorities. The difference between the upper and lower bounds of the largest and smallest classes can be used to approximate the range for grouped data. The following formula can be used to get the relative measure that corresponds to the range or coefficient of range.

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

Variance

Variance measures how "spread out" the information focuses on the typical value. A high difference shows a more widespread, for certain scores a lot higher and some much lower than the mean. The last fluctuation reflects the normal of these squared deviations, giving a proportion of how much the information has increased. Variance is frequently matched with standard deviation (SD), which is basically the square foundation of difference.

Standard Deviation

The most popular and significant measure of variance is the standard deviation. Signs are not taken into account while calculating the average deviation. This issue is settled by the standard deviation, which squares the variances to make them all positive. The Greek letter an in lower case, which may alternatively be read as sigma, is typically used to represent the standard deviation, sometimes referred to as the root mean square deviation. This can be expressed in symbols as

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

Variance is defined as the standard deviation squared. Both the standard deviation and variance increase with the square of the data. Most importantly, it can be easily compared to other standard deviations, and the variability increases with the standard deviation. To understand the formula for grouped data, consider the following data, Which relate to the profits of 100 companies.

Profits(Rs. lakhs)	No of companies	Profits (Rs. lakhs)	No of companies
8-10	8	14-16	30
10-12	12	16-18	20
12-14	20	18-20	10

4.5 Applications of Variation and Skewness in Business Analysis

Grasping the spread (variation) and shape (skewness) of that information is necessary for separating significant experiences and settling on informed choices. This is the way variation and skewness become possibly the most important factor:

Variation

1) **Grasping Client Conduct:** Client conduct can fluctuate fundamentally. By breaking down historical information, organisations can understand fluctuations in money management. A high fluctuation could indicate a client base with different requirements and spending designs.
2) **Stock Administration:** Compelling stock administration depends on grasping interest variances. By breaking down information on past deals, organisations can survey the fluctuation of every day, week-by-week, or occasional deals. High fluctuation popularity requires a more adaptable stock methodology to stay away from stock-outs or surplus stock-holding costs.

Skewness

1) Market Investigation: Market patterns don't necessarily follow a perfectly straight line. By dissecting market information (e.g., lodging costs), organisations can survey skewness. A right angle, with a more drawn-out tail towards higher qualities, could show a market with a couple of top-of-the-line exclusions and a greater part of mid-range values.
2) Deals Execution: Marketing projections frequently show skewness. By investigating deal information, organisations can recognise a positive slant, where few sales reps reliably outflank the greater part.

- **Knowledge Check Activity -2**

  State True or False

  1. To have a better understanding of the data scattering, we measure the two values with differing units using relative measures of dispersion. (True)
  2. Standard Dispersion refers to the dispersion metrics that are measured and expressed in the actual data units. (False)
  3. The arithmetic mean of the variation between the values and their mean is known as the mean deviation. (True)

4. A collection of data's dispersion doesn't reveal its consistency or variability. (False)

- **Outcome-Based Activity -2**

  Consider yourself a teacher who has given each of your pupils a project. After grading the projects, you discovered that although some students' ratings were quite close to one another, others' results ranged over a considerably larger range. How would you characterise the distribution of student marks on this project using the idea of dispersion or spread?

4.6 Summary

- Measuring variability highlights how representative an average is of the whole data, which helps establish its reliability.
- The fact that the quartile deviation is the sole variability measure appropriate for open-ended distributions is another benefit of this measurement.
- It is the sole metric with the required mathematical characteristics (such as combined standard deviation) to be helpful in difficult statistical analysis.

4.7 Keywords

- **Measure of Variation** – The spread or scattering of the individual values around the centre value is described by a measure of variation
- **Range**- The difference between a collection of data's greatest (numerically largest) and lowest (numerically smallest) values is known as the range.
- **Quartile Deviation**—The difference between the third and first quartiles is averaged to get the quartile deviation, which is sometimes referred to as the semi-interquartile range.
- **Coefficient of variation** – It is a percentage representation of the standard deviation to mean ratio.
- **Skewness** – The absence of symmetry in distribution is referred to as skewness.
- **Dispersion** – In statistics, dispersion is a term used to characterise how dispersed or distributed the data is around an average value.

4.8 Self-Assessment Questions

1) What are the different measures of variation?
2) What do you mean by relative Skewness?
3) What are the different types of dispersion?
4) What are the major differences between measures of dispersion and central tendency?

4.9 References/ Reference Reading

- Gupta, S.P., and M.P. Gupta. Business Statistics. Sultan Chand & Sons, 2020.
- Kothari, C.R., and Gaurav Garg. Research Methodology: Methods and Techniques. New Age International Publishers, 2019.
- Sharma, J.K. Business Statistics: Problems and Solutions. Pearson Education, 2021.
- Das, N.G. Statistical Methods. Tata McGraw-Hill Education, 2018.
- Levin, Richard I., and David S. Rubin. Statistics for Management. Pearson Education, 2017.

Unit 5: Introduction to Correlation and Regression Analysis

Learning outcomes

- Students will be able to define the meaning of correlation and regression
- Students will be able to evaluate the significance of correlation and regression
- Students will be able to analyse the importance of a simple linear regression model
- Students will be able to understand the assumptions of a simple linear regression model
- Students will be able to remember the applications of correlation and regression in business.

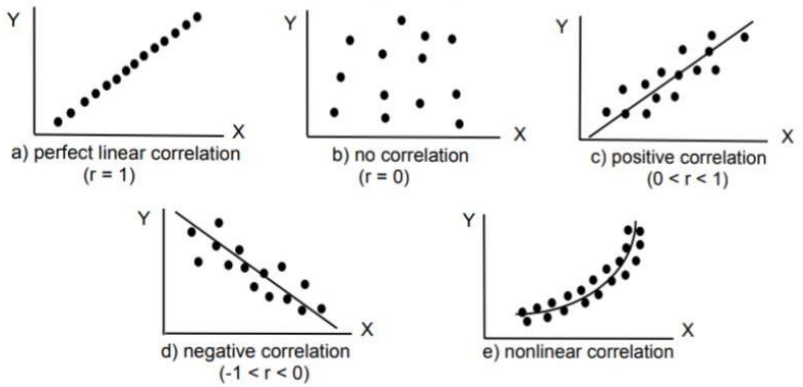
Structure

5.1 Definition and Significance of Correlation and Regression

- **Meaning of Correlation**

A mathematical depiction of the direct, or "straight-line," connection between two constant factors, X and Y, is given by relationship. After the mathematician who initially characterised it, the connection coefficient or "r esteem" that results is known as the Pearson item second connection coefficient. Though Y is indicated as the ward or reaction variable, X is referred to as the free or logical variable. The connection coefficient has the significant advantage of having the option to think about any two factors, no matter what their units are, as it is autonomous of X and Y's units.

Drawing the observations in a "scattergram" or "disperse plot" is a vital starting move toward deciding a connection coefficient since it permits you to outwardly evaluate the information for potential connections or the presence of exception values. A smooth curve can ordinarily be seen through the information, allowing one to decide the sort of relationship that is there. Plotting the reliant variable on the Y-pivot and the free factor on the X-hub is a standard technique. Since Y rises (or diminishes) by a similar sum as X when X shifts, we might reason that X is completely due to the fault of Y's change.



There is a significant link on the off chance that the information focuses on an oval structure, and the r esteem falls somewhere in the range of 0 and 1. At the point when the dependent variable ascents pairs with the free factor, there is a positive relationship. At the point when the reliant variable slopes while the free factor falls, or the other way around, there is a negative connection. A critical yet nonlinear affiliation could slip through the cracks if a scattergram of the information isn’t displayed before it is set in stone.

How much change in the reliant variable Y that is believed to be made sense of by an adjustment of the free factor X is communicated as the square of the r esteem,

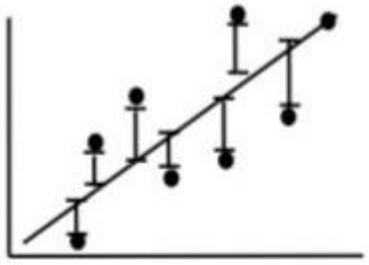
sometimes denoted as the coefficient of assurance. For example, assuming two factors have a r worth of 0.40, the coefficient of assurance is 0.16, implying that an adjustment of X can represent 16% of the variety in Y. The more changes in the autonomous variable affect the reliant variable, the higher the connection coefficient and the greater the coefficient of assurance.

Meaning of Regression

Regression analysis quantitatively reveals the dependence of the Y variable on the X variable and fosters a condition which can be used to predict any value of Y for any value of X. Appears differently in relation to association; it is more point-by-point and offers more information. It is often suggested to be a clear backslide, direct backslide, or at least a square backslide.

Y= a+bX

The "least squares" approach, which limits the amount of the squared vertical distances between the genuine and expected upsides of Y, is utilised to fit the relapse line. Revert examination yields one more adjusting measurement through the decline condition, slant, and block: the standard mistake of the incline.



The standard mistake of the incline is a measure of how intently the deliberate slant approaches the genuine slant, much as the standard blunder of the mean. It is a gauge of how intently the example mean approximates the populace mean. It is obtained from the residuals' standard deviation and filled in to check the relapse line's "decency of fit" to the information. Every information point's upward distance is known as a remaining.

- **Significance of correlation**

1) It guides in sorting out how firmly the two factors in an independent figure are related.
2) Correlation is a helpful device in business direction. Making conjectures about how to utilize the relationship assists in reducing vulnerability. This is the case since connection-based estimates are most frequently exact and firmly aligned with the real world.

- **Significance of regression**

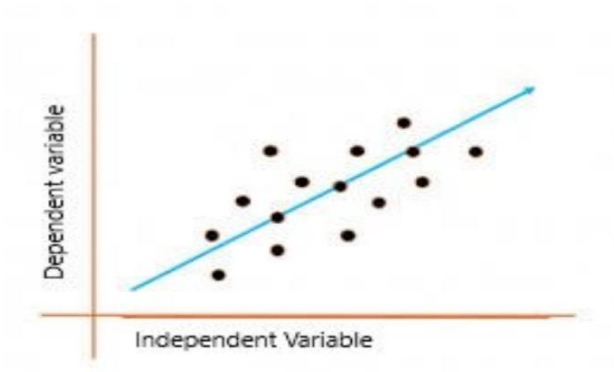
1) A fundamental component of statistics, regression analysis is extremely important for comprehending correlations between variables. This is the reason it's crucial:
2) Revealing Relationships: Regression aids in calculating the relationship between one or more independent variables (caused by the changes) and a dependent variable (influenced by the changes). It clarifies the intensity and direction of this relationship.
3) Predictive Power: Regression enables us to forecast future results by revealing the relationship between variables.
4) Explanatory Tool: Regression aids in explaining relationships in addition to displaying them. By looking at the coefficients linked to it, we may determine each independent variable's specific contribution to the dependent variable.

5.2 Simple linear regression model

- **Meaning of simple linear regression**

A fundamental direct relapse consists of one free factor and one dependent variable. The line of best fit addresses the connection between the factors whose slant and capture are assessed by the model. The block shows the normal worth of the reliant variable when the free factor is zero. However, the slant shows the adjustment of the reliant variable for every unit change in the autonomous variable.

The graph above shows a linear relationship between the result (y) and predictor (X) variables.



The blue line denotes the best-fit straight line. We try to draw a line that most closely fits the provided data points.

- **Importance of simple linear regression model**

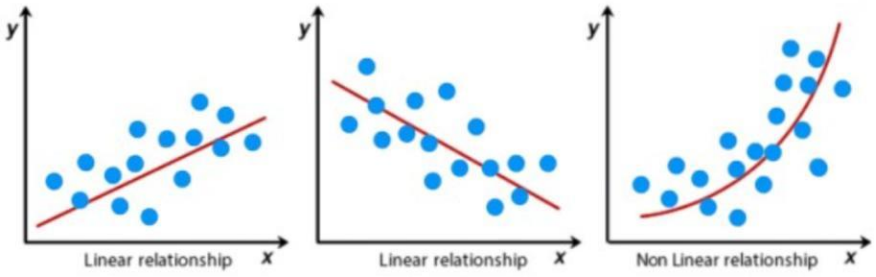
There are a few motivations behind why straight relapse is significant:

1) Simplicity and interpretability: It's a sensibly straightforward plan to understand and incorporate. The resultant fundamental direct relapse model is a straightforward condition that outlines the connection between one variable and another.
2) Prediction: Utilising current information, straight relapse empowers you to figure out future qualities.

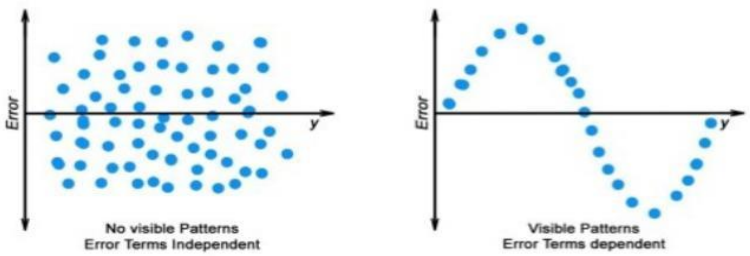
- **Assumptions of simple linear regression**

As a parametric technique, regression relies on assumptions about the data to be analysed. Regression analysis requires the following assumptions to be validated in order to be successful.

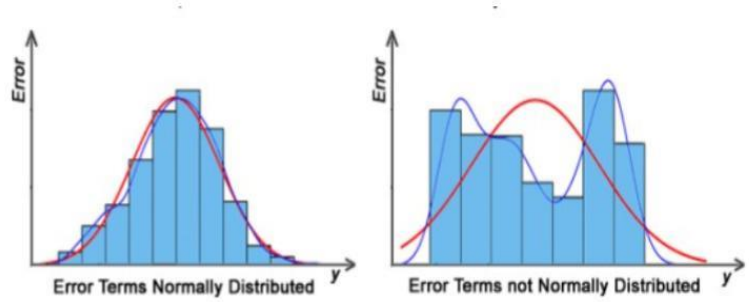
1) Linearity of residuals: The dependent variable and the independent variable(s) must have a linear relationship.



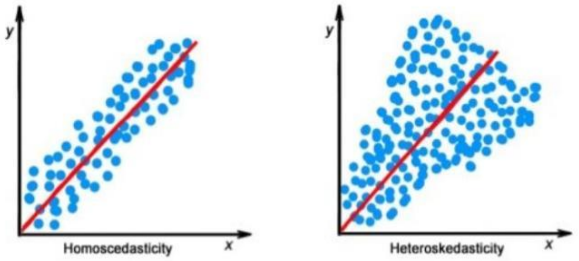
2) The independence of residuals refers to the absence of dependency between the error terms, unlike with time-series data, where the subsequent value is contingent upon the preceding one. The rest of the terms shouldn't be correlated with one another. Autocorrelation is the lack of this phenomenon. The incorrect phrases shouldn't exhibit any detectable patterns.



3) Normal distribution of residuals: A residual mean that is equal to or nearly equal to zero should be distributed normally. This is done to determine if the chosen line is, in fact, the line of best fit.



4) The error terms must have a consistent variance in order for the residuals to have equal variance. We refer to this as homoscedasticity. Heteroscedasticity is the state in which the error terms exhibit non-constant variation. Non-constant variance typically occurs when excessive leverage levels or outliers are present.



- **Knowledge Check 1**

  Fill in the Blanks

  1. A numerical representation of the linear, or "straight-line," relationship between two continuous variables, X and Y, is given by_____. (correlation)
  2. _____ analysis quantitatively defines the dependency of the Y variable on the X variable and constructs an equation that can be used to predict any value of Y for any value of X. (Regression)
  3. The _____ shows the expected value of the dependent variable when the independent variable is zero. (Intercept)

4. A _____ mean that is equal to or nearly equal to zero should be distributed normally. (Residual)

- **Outcome Based Activity -1**

  Assume you collected information on your peers' exam results (out of 100) and hours spent studying. To graphically depict this data, make a scatter plot. Justify your expectations for a positive or negative linear relationship between exam scores and studying.

5.3 Coefficient of Regression

- **Meaning of coefficient of regression**

The meaning of regression coefficients is assessments of specific qualities used to portray the association between a reaction and an indicator variable.

Properties of the regression coefficient

1) 'b' is regularly used to demonstrate it.
2) It is expressed with regard to a special information unit.
3) Two upsides of the regression coefficient are delivered, assuming that there are two factors, suppose x and y. At the point when x is free, and y is reliant, one will be procured; the other will happen when y is viewed as autonomous and x is reliant. Byx and bxy mirror the revert coefficients of y on x and x on y individually.
4) The indication of the relapse coefficients should match the two of them. Bxy will moreover be positive in the event that byx is positive, and the inverse is additionally obvious.

5.4 Applications of Correlation and Regression in Business

In the regularly developing world of business, understanding connections between factors is significant for settling on informed choices.

Correlation:

1) Showcasing and Deals: Connection investigation recognizes possible connections between promoting endeavours and deals with results.
2) Client Division: Understanding client inclinations is fundamental for designated advertising. Connection investigation can assist with recognising connections between client socioeconomics (age, pay) and buy history.

Regression

1) Estimating Procedures: Regression analysis goes past essentially distinguishing a connection; it permits us to evaluate the connection between factors. Envision examining the connection between items, including screen size) and cell phone costs.
2) Request Anticipating: Precise interest estimating is basic for stock administration. Relapse investigation can be utilised to display the connection between verifiable deals information and elements like irregularity or promoting efforts.

- **Knowledge Check 2**

  State True or False

  1. Regression Coefficient is stated in terms of a unique data unit. (True)
  2. Constant variance typically occurs when excessive leverage levels or outliers are present. (False)
  3. When using a known variable to forecast the value of an unknown variable, linear regression's regression coefficients come in handy. (True)
  4. Regression coefficients will be more than unity if one is bigger than unity. (False)

- **Outcome-Based Activity -2**

  Assume that you studied the association between exam scores (dependent variable) and study hours (independent variable). After doing a linear regression analysis, the studying hours variable had a coefficient of 2. Briefly explain what, in this case, a coefficient of 2 indicates regarding the correlation between study time and test results.

5.5 Summary

- A numerical representation of the linear, or "straight-line," relationship between two continuous variables, X and Y, is given by correlation.
- A smooth curve can usually be seen through the data, allowing one to determine the kind of relationship that is there.
- When the dependent variable rises in tandem with the independent variable, there is a positive correlation.

- Rank correlations can also be employed when comparing ordinal or discrete variables (like live vs. die or disease against no disease) with continuous variables (like cardiac output).
- Regression analysis quantitatively defines the dependency of the Y variable on the X variable. It constructs an equation which can be used to predict any value of Y for any value of X.
- Correlation is a useful tool in commercial decision-making.
- When the individual groups are not statistically associated, the aggregated observations may seem to be.

5.6 Keywords

- **Correlation**: A numerical representation of the linear, or "straight-line," relationship between two continuous variables, X and Y, is given by correlation.
- **Simple linear regression:** This type of linear regression is known as simple linear regression if there is just one input variable, X (the independent variable).
- **Regression Coefficient:** The definition of regression coefficients is estimations of certain unknown characteristics used to characterise the connection between a response and a predictor variable.

5.7 Self-Assessment Questions

1. What do you mean by correlation?
2. What is the significance of regression analysis?
3. What is the importance of the Simple linear regression model?
4. What are the properties of the regression coefficient?
5. What are the assumptions of a simple linear regression model?

5.8 References/ Reference Reading

- Das, Narayan, and Sanjay Srivastava. *Business Statistics: Theory and Applications*. McGraw Hill Education, 2022.
- Sharma, J. K. *Business Statistics*. Pearson Education India, 2021.
- Gupta, S. P., and Archana Gupta. *Statistical Methods*. Sultan Chand & Sons, 2020.
- Anderson, David R., et al. *Statistics for Business & Economics*. Cengage Learning, 2021.

- Keller, Gerald. *Statistics for Management and Economics*. Cengage Learning, 2022.

Unit 6: Methods of Correlation Analysis

Learning Outcomes

- Students will be able to define the interpretation of the scatter diagram
- Students will be able to analyse the methods of calculating Karl Pearson's Coefficient of Correlation
- Students will be able to understand the merits and Demerits of the rank correlation coefficient
- Students will be able to remember the advantages and disadvantages of the least squares method
- Students will be able to remember the meaning of standard error of estimates.

Structure

6.9 Self-Assessment Questions

6.10 References/ Reference Reading

6.1 Scatter Diagram

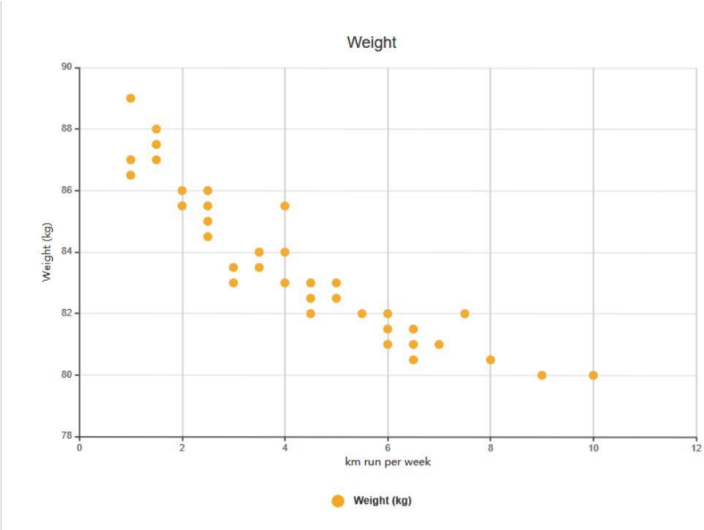
- **Meaning of scatter diagram**

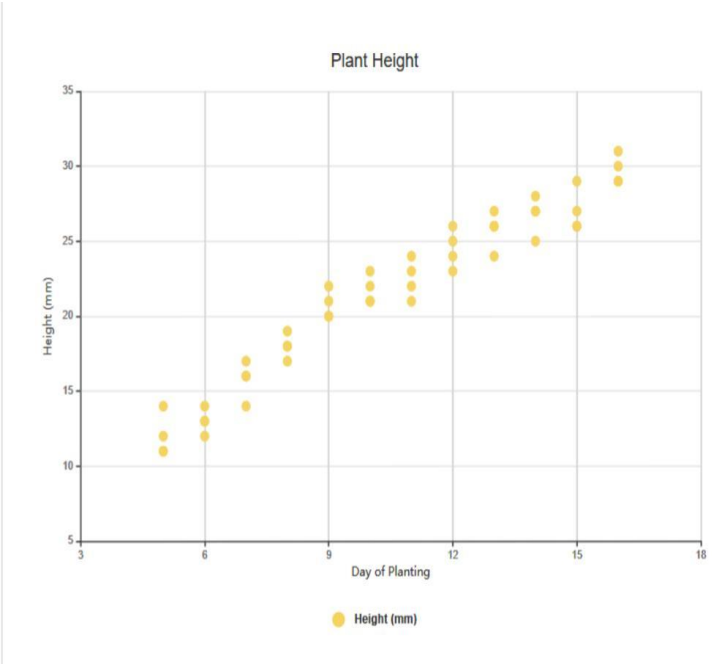
The scatter diagram approach is a typical graphic aid for determining the connection between two variables in statistics and business. On a two-dimensional graph, these two variables are displayed along the X and Y axes, and the pattern shows the relationship between them. Scatter diagram analysis is the examination of a two-variable graphical representation using a scatter diagram.

- **Types of scatter diagram**

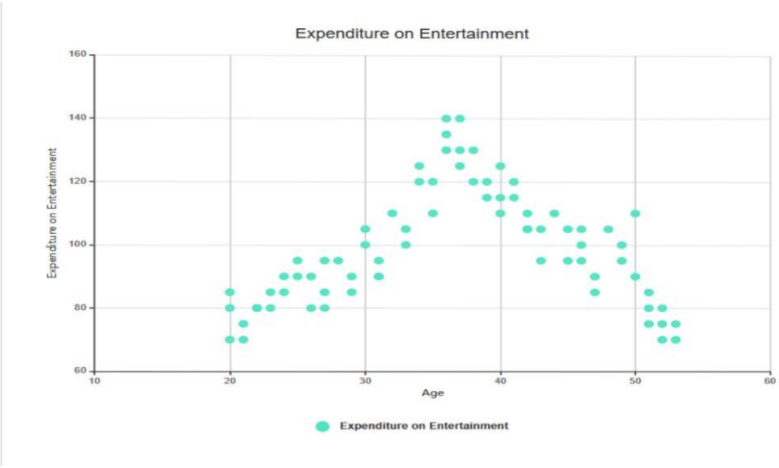
To help the students, it is important to make sense of the dispersed outline utilising models while appreciating its many structures remarkably; while there are a few portrayals that each infer different types of affiliation, the most pervasive and critical ones are examined here.

1) Perfect Positive correlation: When every one of the plots focuses on a chart and is shown on a diagram, the connection is supposed to be great.

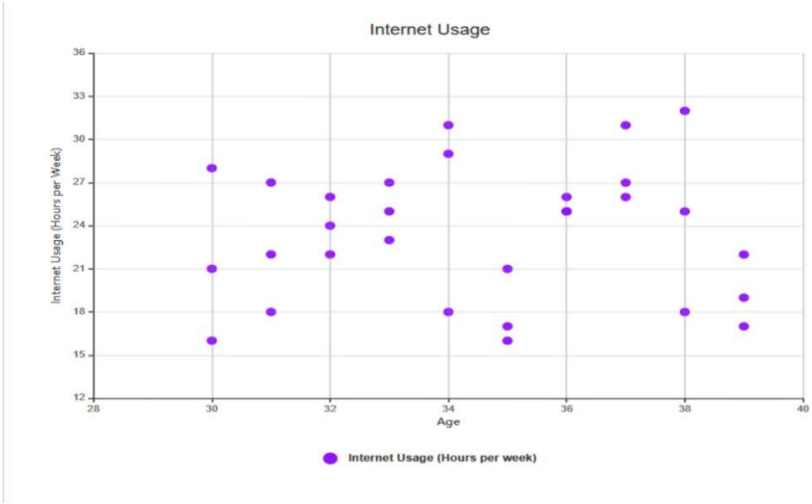


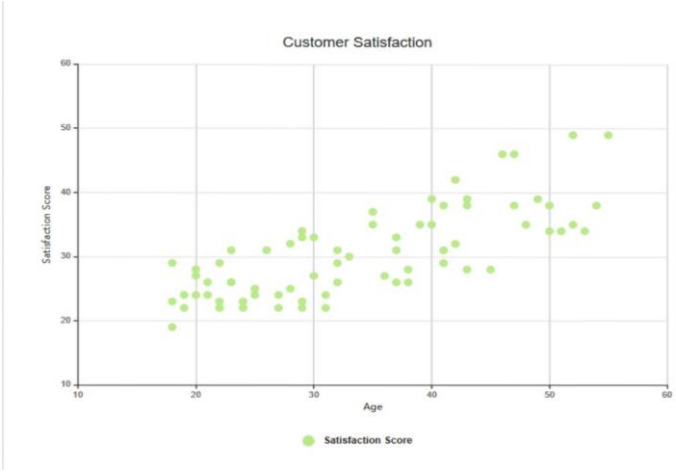


2) Perfect Negative Correlation: An ideal negative connection is complementary to the previous kind in dispersed outline models. Here, as well, every single plotted point is completely on a straight line.
3) Low Degree of Positive Correlation: Plotted focuses are dissipated when there is a low level of connection, either positive or negative. These portrayals' graphs look like the portrayal displayed beneath.



4) Low Degree of Negative Correlation: A scatter point graph, such as the one shown above, illustrates low degrees of negative correlation.





5) No Correlation: Although the description of a scattering diagram focuses on determining the correlation between variables, students also need to be aware that the representations may be dispersed and incoherent.

6.2 Karl Pearson's Coefficient of Correlation

- **Meaning of Karl Pearson's Coefficient of Correlation**

Karl Pearson was quick to give a numerical strategy for deciding the level of relationship between two factors in 1890. The Item Second Relationship or Basic Connection Coefficient are different names for Karl Pearson's Coefficient of Relationship. This one is the most widely involved and popular strategy for ascertaining the coefficient of relationship. It is addressed by the image "r," where r addresses an unadulterated whole number (i.e., it has no unit).

Karl Pearson states that the computation of the coefficient of relationship includes separating the absolute results of the deviations from the particular means by the number of matches and standard deviations.

- **Methods of Calculating Karl Pearson's Coefficient of Correlation**

$$r = \frac{\sum xy}{N \times \sigma_x \times \sigma_y}$$

Where,

N = Number of Pair of Observations

x = Deviation of X series from Mean $(X - \bar{X})$

y = Deviation of Y series from Mean $(Y - \bar{Y})$

σ_x = Standard Deviation of X series $(\sqrt{\frac{\sum x^2}{N}})$

σ_y = Standard Deviation of Y series $(\sqrt{\frac{\sum y^2}{N}})$

r = Coefficient of Correlation

1) Actual mean method – The means engaged with the computation of the coefficient of connection by utilising the Actual Mean Method are
   A) Finding the mean of the two series — suppose X and Y — is the initial step.
   B) Now, consider the X series' flight and utilize x to demonstrate the varieties.
   C) Square the x deviations to view as the aggregate
   D) Consider the Y series' flight and demonstrate the y-contrasts.
   E) Square the y-distances to track down the aggregate.
   F) To track down the aggregate, increase the comparing deviations of Series X and Y.
   G) Now, ascertain the coefficient of connection utilising the equation below:

$$r = \frac{\sum xy}{\sqrt{\sum x^2 \times \sum y^2}}$$

2) Direct method –: The means engaged with the computation of the coefficient of connection by utilising the Direct Method are:
   A) The completion of Series X should be resolved first.
   B) Subsequently, get the Series Y total.
   C) Compute the number of the X Series values by establishing them by square.
   D) Square the Y Series values and track down their total.

E) Calculate the number of qualities from Series X and Y by increasing them together.

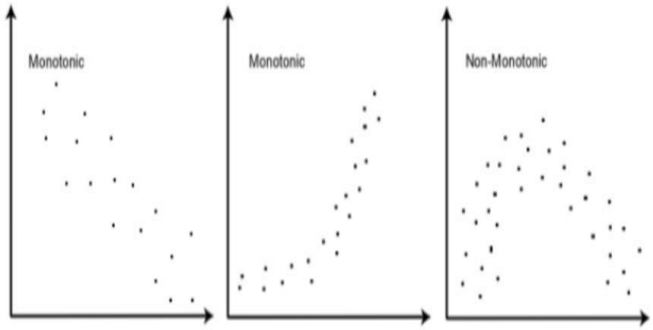
F) Now, work out the coefficient of the relationship utilising the equation below:

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

6.3 Rank Correlation

- **Meaning of rank correlation**

The nonparametric Spearman's Connection Coefficient, signified as ρ or rR, evaluates the degree and direction of the relationship between the two positioning factors. It lays out on the off chance that there is a monotonic part to the connection between two constant or requested factors or how monotonic a relationship is.



- **Merits and Demerits of a Rank correlation coefficient**

Merits

1. Spearman's position relationship coefficient is direct to process and understand.
2. It could be interpreted similarly to Karl Pearson's connection coefficient.
3. Rank relationship coefficient is the non-parametric form of Karl Pearson's item second connection coefficient.

Demerits

4. For a bivariate repetition conveyance, the item's second relationship coefficient might be processed; however, the position connection coefficient can't be processed.

5. This estimation demands a ton of investment if n > 30.

- **Knowledge Check 1**

  Fill in the Blanks

  1. A typical visual help for determining the connection between two variables in statistics and business is the _____ approach. (Scatter diagram)
  2. When all of the plotted points on a scatter diagram are shown on a graph, the correlation is said to be _____. (Perfect)
  3. Karl Pearson was the first to provide a mathematical method for determining the degree of association between two variables in _____. (1890)
  4. Plotted points are _____ when there is a low degree of correlation, either positive or negative. (Scattered)

- **Outcome Based Activity -1**

  Assume you have information on the movie reviews of a group of ten films from two distinct critics. Determine the Spearman rank correlation coefficient to analyse the data. Even if they don't give the movie precisely the same ratings, do you believe the two critics generally agree on the relative rankings of the movie? Give a brief explanation of your response.

6.4 Method of Least Squares Method

- **Meaning of the least squares method**

Regression analysis in the form of the least squares approach gives the general explanation for where the line of greatest fit should be placed among the data points under study. The first step is to plot a set of data points using two variables along the x- and y-axes on a graph. This is a useful tool for traders and analysts to identify bullish and negative market patterns as well as possible trading opportunities. The most popular form of this approach is often called "simple" or "linear". It attempts to draw a straight line representing the difference between the observed value and the value predicted by that model minus the sum of the squares of the errors introduced by the solutions of the relevant equations.

Advantages and disadvantages of the least squares method

Advantages

- The fact that this technology is easy to use and understand is one of its main advantages. This is because it shows the best relationship between two variables by using only two—one specified on the x- and y-axes
- The least-squares method enables analysts and investors to predict future trends in stocks and finance by analysing historical performance. It can be used as a decision-making tool.

Disadvantages

- The data utilised in the least squares approach is its main drawback. It is limited to illustrating the correlation between two variables. It does not consider any other as a result. Additionally, the findings are biased if there are any outliers.
- The need for the data to be dispersed equally presents another issue with this approach. The findings might not be truthful if this isn't the case.

6.5 Standard error of estimates

With regards to regression analysis, the standard error of estimates (SEE) turns into a significant measure to assess a fitted model's viability. It estimates the typical distinction between the qualities in the informational index compared to the noticed qualities and the expected qualities delivered by the relapse line. Basically, the SEE addresses the vulnerability level characteristic of the model's estimates.

Think about a circumstance in which the association between a free factor (X) and a reliant variable (Y) is demonstrated utilising direct relapse. To distinguish the basic pattern in the scatterplot of the data of interest, the relapse line is utilised as a delegate line. The SEE decides the typical distance on the relapse line between every data of interest (Y) and its related projected esteem (). The SEE, which is addressed in similar units as the reliant variable (Y), is acquired by working out the square base of the mean squared deviations from the normal qualities.

6.6 Multiple correlation and partial correlation

In the branch of measurements, connection examination permits us to evaluate the connections between factors. However, what happens while we're managing a large number of variables that could impact a solitary result? This is where numerous connections and halfway relationships move toward, offering a more nuanced comprehension of intricate connections.

Multiple Correlation

1) Promoting Efforts: Current showcasing efforts frequently include a blend of techniques like virtual entertainment publicising, site enhancement, and email advertising. Various relationship investigations can survey the consolidated effect of these factors on a solitary result, like site changes.
2) Monetary Examination: Stock costs are impacted by different elements, in addition to a solitary organisation's exhibition. Numerous relationship examinations can assess the joint impact of monetary pointers (financing costs), industry execution (contender stock costs), and company-explicit variables (profit provides details regarding) a stock's cost development.

Partial correlation

While various relationships reveal the combined impact of numerous factors, a halfway connection permits us to separate the effect of one variable while measurably controlling for the impacts of others. Visualise the situation of dissecting the connection between web composition (X1) and online deals (Y) while likewise thinking about the effect of promoting endeavours (X2).

The fractional connection between web architecture (X1) and online deals (Y), controlling for advertising endeavours (X2): This investigation separates the exceptional impact of web composition on deals, barring the effect of showcasing efforts.

- **Knowledge Check 2**

  State True or False

  1. SEE is a useful tool for assessing how generalisable the model is and how well it can forecast future events based on the established link between the variables. (True)
  2. Regression analysis doesn't assume that the errors in the independent variable are small or zero. (False)
  3. Spearman's rank correlation coefficient is straightforward to compute and comprehend. (True)
  4. Variance analysis in the form of the least squares approach gives the general explanation for where the line of greatest fit should be placed among the data points under study. (False)

- **Outcome-Based Activity -2**

  Your job is to fit a straight line to a dataset that shows the relationship between exam scores (dependent variable) and study hours (independent variable). Explain the least squares approach to this situation. How would using this strategy assist you in figuring out which line best fits the data?

6.7 Summary

- Scatter diagram analysis is the examination of a two-variable graphical representation using a scatter diagram.
- When all of the plotted points on a scatter diagram are shown on a graph, the correlation is said to be perfect.
- Plot the dependent variable on the y-axis and the independent variable on the x-axis of a graph.
- If a pattern emerges, notice it. If the dots form a clear curve or line. It suggests a correlation between the variables.
- The least-squares approach is a particularly useful curve-fitting technique.
- This is due to the fact that it highlights the optimal relationship between the two variables using just two—one shown along the x- and y-axes.

6.8 Keywords

- **Scatter diagram analysis** – Scatter diagram analysis is the examination of a two-variable graphical representation using a scatter diagram.
- **Perfect positive correlation** – When all of the plotted points on a scatter diagram are shown on a graph, the correlation is said to be perfect.
- **Least squares method**—The least squares approach to regression analysis gives a general explanation for where the line of greatest fit should be placed among the data points under study.
- **Standard error of estimates**—This measures the average difference between the values in the data set that correspond to the observed values and the anticipated values produced by the regression line.

6.9 Self-Assessment Questions

1. What are the types of scatter diagrams?

2. What do you mean by Karl Pearson's Coefficient of Correlation?
3. What are the merits and Demerits of the rank correlation coefficient?
4. What are the limitations of the least squares method?
5. What do you mean by Standard error of estimates?

6.10 References/ Reference Reading

- Gupta, S. P., and M. P. Gupta. *Business Statistics*. Sultan Chand & Sons, 2020.
- Das, N.G. *Statistical Methods*. McGraw Hill Education, 2019.
- Hogg, Robert V., and Elliot A. Tanis. *Probability and Statistical Inference*. Pearson, 2018.
- Anderson, David R., et al. *Statistics for Business and Economics*. Cengage Learning, 2019.
- Srivastava, T. N., and Shailaja Rego. *Statistics for Management*. Tata McGraw-Hill, 2017.

Unit 7: Time Series Analysis

Learning outcomes

- Students will be able to understand the types of data in time series analysis.
- Students will be able to learn about the utility of time series analysis.
- Students will be able to know the techniques of time series analysis.
- Students will be able to learn about the components of time series analysis.
- Students will be able to understand the steps involved in the analysis of time series.

Structure

7.1 Introduction to Time Series Analysis

- **Meaning of Time Series Analysis**

In information science, measurements, and examination, time series examination is fundamental. The key objective of time series examination is to look at and assess a progression of information focuses that have been caught or accumulated at customary spans. Rather than cross-sectional information, which keeps an independent moment, time series information is fundamentally powerful, changing all through a large number of sequential groupings from remarkably short to extremely lengthy. Tracking down hidden structures in the information, like cycles, patterns, and occasional variances, relies vigorously upon this sort of exploration.

- **Types of Data in Time Series Analysis**

Understanding the type of information you're working with is regularly the most vital phase in starting a time series examination. Three fundamental classes include this order: Time Series Information, Cross-Sectional Information, and Pooled Information. Each sort has unmistakable attributes that immediately affect the resulting demonstration and examination.

i) Time series information is comprised of perceptions made at different moments. Dissecting different worldly examples, including cycles and trends, is planned.
ii) Cross-sectional information: Comprises of data accumulated at one explicit moment. They were accommodating in understanding associations or differences between numerous components or classifications at a specific time.

7.2 Utility of Time Series Analysis

- **Types of Time Series Analysis**

1) Investigative study

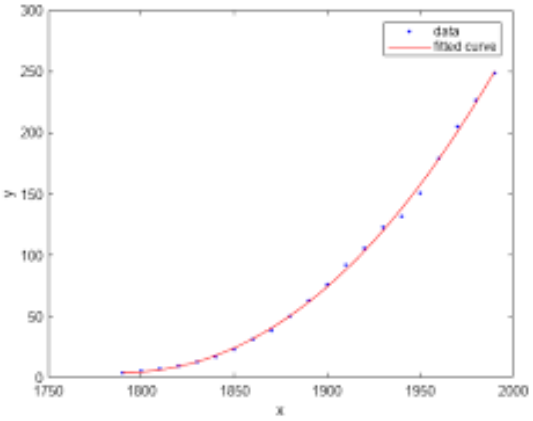
While introducing what you see in a specific time series and clarifying why you see it, exploratory examination is helpful. Breaking down the information into patterns, irregularity, cyclicity, and abnormalities is fundamentally what it includes.

2) Fitting curves

The quantity of data of interest in a period series can continuously be set definitively since it is a distinct collection.

We want to add a continual put to our information — a bend — together to have the option to answer this inquiry. There are a few strategies for doing this, including relapse and interjection. The previous is, for the most part, supportive of anticipating missing

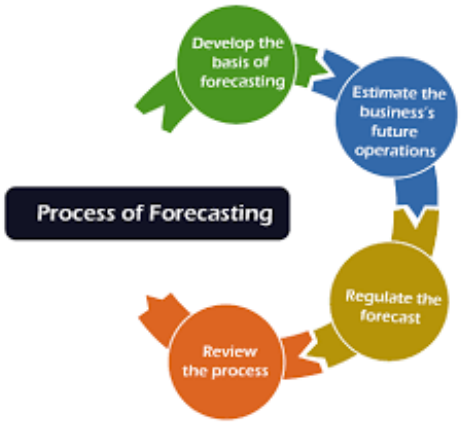
data of interest and matches precisely for specific portions of the given time series. The last option, then again, is a "best-fit" bend where you should gauge the state of the capability to be fitted (direct, for instance) and afterwards change the boundaries until your best-fit rules are fulfilled.



A "best-fit" situation fluctuates as per the test and planned outcome. Furthermore, you might utilize relapse examination to track down the best-fit capability boundaries.

3) Forecasting

Summing up from test to entire is the course of factual derivation. It might very well be completed step by step in time-series information, preparing for estimating or future expectations, from straightforward relapse model extrapolation to additional mind-boggling strategies utilising AI and complex recreations.



- **Utility of Time Series Analysis**

Time series examination is crucial for financial specialists and money managers, as well as researchers, geologists, scholars, and scientists.

1) It works with an understanding of earlier activities. By surveying information throughout some stretch of time, one may promptly grasp what changes have

occurred previously. Such examination will be exceptionally beneficial in creating future ways of behaving.

2) It is helpful for sorting out forthcoming tasks. It isn't easy to make future arrangements without projecting the occasions and the connections they will include. The improvement of measurable devices has made it possible to break down time series so that it is possible to find the elements that moulded their structure.

- **Knowledge Check 1**

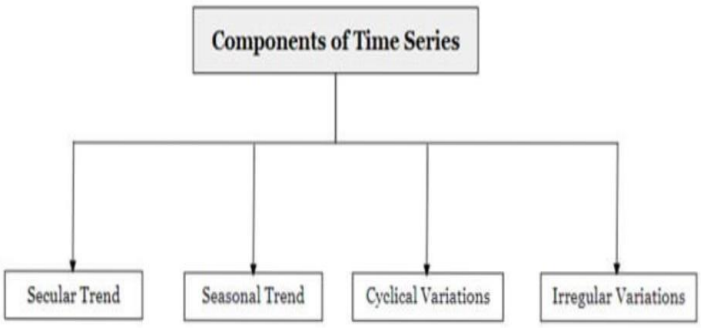
  Fill in the blanks

  1. The fundamental goal of _____ is to examine and evaluate a succession of data points that have been captured or gathered at regular intervals. (Time series analysis)
  2. When presenting what you observe in a particular time series and explaining why you perceive it, _____ is useful (Exploratory analysis)
  3. The process of classifying an outcome variable or classes using time-series data is known as time-series _____. (Classification)
  4. The source of _____ may be examined, and the performance can be compared to the predicted performance. (Variance)

- **Outcome-Based Activity 1**

  A retail store manager is reviewing last year's sales information. The data shows daily sales numbers. Describe the potential benefits of time series analysis for management and the insights that it may provide.

7.3 Components of Time Series

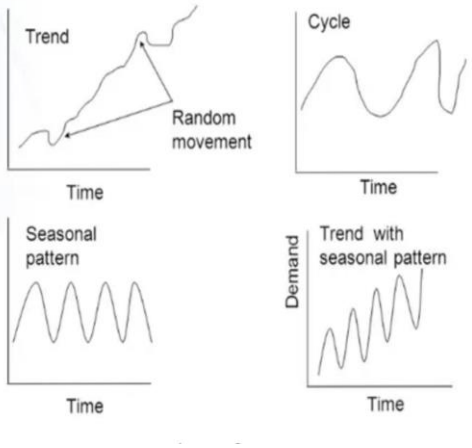


1) Secular Trend

It shows the getting-through design found in the grouping of information that was caught. It could be moving vertically or descending, connoting the course representing things to come. Regardless of being generally perceived as the normal propensity of any perspective, the pattern can vary in specific regions, switching back and forth between up and descending movement.

2) Irregular variations

It compares inescapable and undeniable varieties. Rather than normal changes or events, irregular changes occur generally and are irrelevant to an example. These varieties are unreasonable and surprising. Occasions, such as artificial and normal disasters, can cause irregular changes.



7.4 Analysis of Time Series

- **Meaning of Analysis of Time Series**

A measurable technique for inspecting and deciphering progressive information gathered over a long time is time series investigation. Contrasted with other measurable examinations, such an information investigation offers bits of knowledge into the hidden examples, patterns, and ways of behaving of a given dataset

- **Importance of Time Series Analysis**

Inside the consistently changing domain of information-driven independent direction, time series examination is a fundamental device. It goes beyond the limitations of static information by focusing on perceptions' ordered arrangement. This time aspect gives admittance to an abundance of previously unseen data.

Time series examination's genuine strength is its ability to close the information gap between previous occasions and future projections. It permits us to effectively influence the "how" representing things to come and understand the "why" behind past events.

- **Examples of analysis of time series**

The accompanying models show the consistent application of time series examination in many fields.

1) Time series examination and stock cost determining in finance- picture what is going on where a stock financial backer wishes to conjecture the cost of that specific stock from here on out. A financial backer can make very educated decisions with respect to buying or discarding stocks in view of expected designs by utilising time series models, like ARIMA or GARCH, to dissect past stock costs.
2) Risk Control in Banking and Money Time series examination is a device utilised by banks to assess market, loaning, and venture risk. To make risk models that would assist them with settling on better loaning and speculation choices, banks could analyse authentic information on advance defaults, financing costs, and monetary pointers.

7.5 Applications of time series analysis in business forecasting

Time series examination is a useful asset utilised in business to determine and foresee future patterns and values in light of verifiable information. Here is a more critical

gander at how time series examination is applied in different business-determining situations:

1. Request Forecasting: Organisations can anticipate future client interest in items and administrations. This assists with stock administration, creation planning, and asset allocation.
2. Monetary Forecasting: Time series investigation is vital for monetary determination. Organisations can foresee future deals in terms of income, costs, and income. This assists with planning, distinguishing monetary dangers, and settling on venture choices.

- **Knowledge Check -2**

  State True or False

  1. Time series analysis is a tool used by banks to evaluate market, lending, and investment risk. (True)
  2. A statistical method for examining and interpreting successive data points gathered over time is time series variance. (False)
  3. One important illustration of a cyclic variation is the business cycle, which indicates that a corporation passes through four stages in its existence. (True)
  4. Seasonal trends are not influenced by artificial customs such as fashion, marriage season, festivals, etc., in addition to natural events. (False)

- **Outcome-Based Activity -2**

  For the last two years, a corporation has been monitoring monthly sales data. Trends, seasonal changes, and erratic fluctuations are probably present in the data. Please explain the process you would use to separate this time series data into its constituent parts. Give a brief explanation of the insights you want to obtain from each part.

7.6 Summary

- The fundamental goal of time series analysis is to examine and evaluate a succession of data points that have been captured or gathered at regular intervals.
- Finding underlying structures in the data, such as cycles, trends, and seasonal fluctuations, depends heavily on this kind of research.

- Extrapolation forecasts are made possible by regression, which yields a function that characterises the best fit to our data even after the last record.
- It may be carried out gradually in time-series data, paving the way for forecasting or future predictions, from simple regression model extrapolation to more complex methods using machine learning and stochastic simulations.

7.7 Keywords

- **Time series analysis:** It aims to represent the underlying structures in the data, taking into consideration trends, seasonal patterns, and autocorrelation.
- **Cross-Sectional Data:** It Consists of information gathered at one specific point in time.
- **Time series classification:** The process of classifying an outcome variable or classes using time series data is known as time series classification.
- **Time series segmentation:** Time-series segmentation involves dividing a time series into segments, each of which represents a distinct state or event.
- **Cyclical Variations:** It symbolises an oscillating pattern of movement that is cyclical.
- **Irregular variations:** It alludes to inevitable variations.

7.8 Self-Assessment Questions

1. What are the types of data in time series analysis?
2. What are the challenges involved in time series analysis?
3. What are the different types of time series analysis?
4. What are the components of time series analysis?
5. What are the steps involved in the analysis of time series?

7.9 References/ Reference Reading

- Bhattacharyya, Dipak, and Ashish Kumar Adhikari. *Statistics: Principles and Applications*. PHI Learning, 2020.
- Brockwell, Peter J., and Richard A. Davis. *Introduction to Time Series and Forecasting*. 3rd ed., Springer, 2016.
- Mahajan, R., and R. M. Badgujar. *Fundamentals of Business Statistics*. Ane Books Pvt. Ltd., 2019.

- Wei, William W. S. *Time Series Analysis: Univariate and Multivariate Methods*. 2nd ed., Pearson, 2019.
- Goon, A. M., M. K. Gupta, and B. Dasgupta. *An Outline of Statistical Theory*. Vol. 2, World Press, 2018.

Unit 8: Measuring Trends in Time Series

Learning outcomes

- Students will be able to define the meaning of the average method

- Students will be able to evaluate the procedure for measuring trends of semi-average method
- Students will be able to analyse types of moving averages
- Students will be able to understand the limitations of the least squares method
- Students will be able to remember the types of exponential smoothing.

Structure

8.1 Semi-average method

- **Meaning of Semi average method**

As the name implies, semi-midpoints are processed this way to decide the pattern values. The midpoints of a series' two parts are denoted as semi-midpoints.

Merits and Demerits of the Average Method

Merits

- The pattern estimation technique is direct.
- It is a goal approach as the pattern of esteem that is obtained by applying it to a specific arrangement of information will not be different for everybody.

Demerits

- This approach can give a straight pattern to the information, no matter what its presence.
- Since we are uncertain of the degree to which the effects of different parts have been taken out, this is just an unpleasant method for assessing patterns.

8.2 Moving Average Method

- **Meaning of moving average method**

Financial bankers and market investigators can use a specialised pointer to determine the course of a pattern. To compute a normal, it includes the monetary security's significant information during a specific period and partitions it out by the whole amount of data of interest. The explanation is named a "moving" normal, and it's refreshed continually to highlight the latest cost information.

- **Merits and Demerits of the Moving Average method**

Merits

- Moving midpoints are helpful for spotting designs. This permits merchants to exploit and appreciate laid-out market designs.
- Because it helps with recognising conceivable cost help, it likewise fills in as an emotionally supportive network.

Demerits

Moving normal computations is a fast and basic technique for spotting designs in monetary instruments. However, they have the accompanying disadvantages:

- Since each product or stock has an alternate cost history, applying general rules to all markets is incomprehensible. A moving average can't portray the continuous changes in their costs.
- Finding a pattern is generally finished to gauge the stock's future qualities. Notwithstanding, sorting out moving midpoints will not give brokers an opportunity to bring in cash in the event that the resource doesn't drift one way or the other.

- **Knowledge Check 1**

Fill in the Blanks

1. The averages of a series' two halves are referred to as _____ (Semi averages)
2. If the amount of data is _____, it is easy to divide the data into two halves by only leaving out the middle year. (Odd)
3. The reason it's termed a "_____" average is that it's updated constantly to reflect the most recent price data. (Moving)
4. A _____ is considered to be in an uptrend when its price is above the moving average line. (Financial instrument)

- **Outcome-Based Activity 1**

  Moving averages are a common tool used by investors to mitigate price swings in the stock market, which can be rather volatile. Let's say you are examining a stock's closing prices over the last 20 days. Determine the data's 5- and 10-day simple moving averages. How can you better comprehend stock price movements with the use of these moving averages?

8.3 Method of Least Squares

- **Meaning of Least Squares Method**

Relapse examination as a least squares approach gives the overall clarification for where the line of most noteworthy fit ought to be placed among the useful pieces of information under study. The initial step is to plot a bunch of information focused utilising two factors along the x- and y-tomahawks on a diagram. This is a helpful device for brokers and experts to recognize bullish and negative market designs and expect to exchange valuable open doors.

- **Advantages and disadvantages of the least squares method**

Advantages

- The way that this innovation is easy to utilise and appreciate is one of its key benefits. This is because it features the ideal connection between the two factors, utilising only two — one displayed along the x-and y-tomahawks.
- With the least squares approach, scientists and financial backers might gauge future examples in the securities exchange and economy by looking at authentic execution. It might consequently be applied as an instrument for independent direction.

Disadvantages

- The information used in the all-squares approach is its primary disadvantage. It is restricted to showing the connection between the two factors. It thinks about no other. Also, the discoveries are one-sided, assuming there are no exceptions.
- The need for the information to be scattered similarly gives another issue to this methodology. The discoveries probably won't be reliable on the off chance that this isn't true.

8.4 Exponential Smoothing

- **Meaning of exponential smoothing**

Time series information is gathered consecutively after some time and shows changes in different spaces like money, financial aspects, promotion, and tasks. Gauging, or foreseeing future time series values, is a basic errand in numerous business and dynamic cycles. Exponential smoothing is generally utilised and is a viable technique for time series gauging that can give exact expectations while being computationally productive. Exponential smoothing is a measurable strategy that uses past perceptions of a period series to predict its future qualities. It is classified as "remarkable" because it dramatically diminishes past perceptions, with later perceptions having higher loads than more seasoned ones.

- **Knowledge Check -2**

  State True or False

  1. By minimising the sum of squares of the errors produced by the solutions to the related equations, it seeks to draw a straight line that represents the difference between the observed value and the value predicted by that model. (True)
  2. Regression analysis doesn’t assume that the errors in the independent variable are small or zero. (False)
  3. The least squares approach, for example, can be used by an analyst to provide a line of best fit that illustrates the possible link between independent and dependent variables. (True)
  4. The data utilised in the least squares approach is its main advantage. (False)

- **Outcome-Based Activity 2**

For a group of software engineers, you have to fit a straight line to represent the link between the independent variable of years of experience and the dependent variable of yearly pay. Explain how to obtain the best-fitting line for this data using the least squares approach.

8.5 Summary

- As the name implies, semi-averages are computed using this approach to determine the trend values.
- The trend is then obtained by plotting the two pairs on graph paper and joining the points with a straight line.
- Since we are unsure of the extent to which the impacts of other components have been removed, this is only a rough means of evaluating the trend.
- By analysing price changes for an asset, analysts can determine support and resistance using the moving average.
- Because it follows the price movement of the underlying asset to provide a signal or indicate the direction of a certain trend, it is referred to as a lagging indicator.
- The other kind of moving average is the exponential moving average (EMA), which responds better to current data points by giving greater weight to the most recent price points.

8.6 Keywords

- **Semi-average method:** As the name implies, semi-averages are computed using this approach to determine the trend values. The averages of a series' two halves are referred to as semi-averages.
- **Moving average method:** In order to calculate an average, it adds up all of the financial security data points during a certain period and divides the total by the entire number of data points.
- **Simple moving average method:** It is calculated by adding up all of the recent data points in a collection and dividing the total by the number of periods.

- **Least squares method:** The least squares approach to regression analysis gives a general explanation for where the line of greatest fit should be placed among the data points under study.

8.7 Self-Assessment Questions

1. What are the merits and Demerits of the average method?
2. What is the procedure for measuring trends by the average method?
3. What are the types of moving averages?
4. What are the major differences between simple moving averages and exponential Moving averages?
5. What are the limitations of the least squares method?

8.8 References/ Reference Reading

- Gupta, S. C., and V. K. Kapoor. *Fundamentals of Mathematical Statistics: A Modern Approach*. 12th ed., Sultan Chand & Sons, 2021.
- Sharma, J. K. *Business Statistics*. 4th ed., Vikas Publishing House, 2018.
- Das, N. G. *Statistical Methods*. 2nd ed., Tata McGraw Hill Education, 2017.
- Montgomery, Douglas C., and Cheryl L. Jennings. *Introduction to Time Series Analysis and Forecasting*. 2nd ed., Wiley, 2015.
- Brockwell, Peter J., and Richard A. Davis. *Introduction to Time Series and Forecasting*. 3rd ed., Springer, 2016.

Unit 9: Introduction to Sampling Theory

Learning Outcomes:

- Students will be able to define the principles of sampling.
- Students will be able to analyse different sampling methods.
- Students will be able to evaluate the differences between two main types of sampling.
- Students will be able to understand the merits and Demerits of different sampling methods.
- Students will be able to remember how to determine sampling in business research.

Structure

9.1 Purpose and principles of sampling

- **Meaning of sampling**

The demonstration of choosing a subset of people or things from a more extensive population to draw deductions about the populace is known as examining measurements. This approach is normally used in circumstances where exploring the whole populace is neither achievable nor possible. As opposed to surveying every person in the gathering, scientists select a delegate test that attempts to imitate the elements of the whole local area.

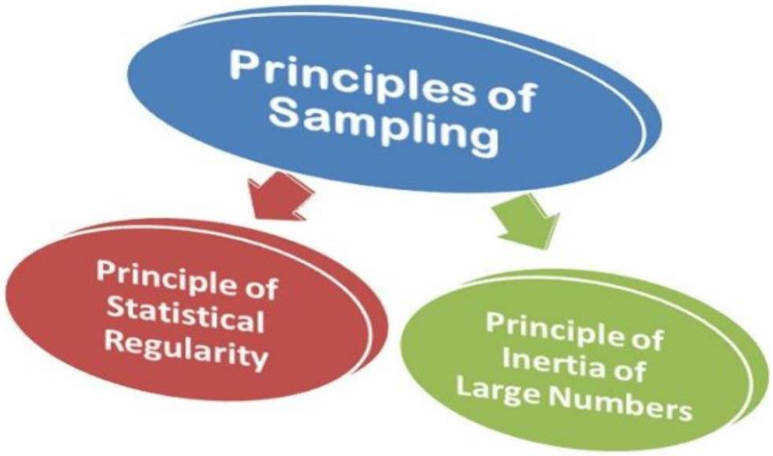
- **Purpose of Sampling**

Via cautiously choosing a subset of information to dissect, testing is a basic system in a factual procedure that permits scientists to make dependable decisions about enormous populaces. This methodology has various essential advantages.

In any case, they tested increment efficiency. Each individual in a population should be inspected, particularly in huge gatherings, which can take a lot of time and resources. Specialists might get data from a sensible number of individuals by picking a measurably significant example, extremely accelerating the information-gathering process.

- **Principles of sampling**

There are two important principles of sampling on which the sampling theory depends: -



1) Principle of "Factual Consistency": The numerical study of likelihood is the cornerstone of the measurable routineness rule. This idea expresses that a critical number of haphazardly picked things from the universe are probably going to have similar qualities as the populace in general.
2) Principle of "Idleness of Enormous Numbers": As indicated by this thought, ends are bound to be more noteworthy than the sample size. The underpinning of this idea is the possibility that numerical numbers have more steady highlights than little numbers. Slight variations in the absolute number of huge numbers don't make any difference.

9.2 Methods of Sampling

- **Meaning of methods of sampling**

The testing approach involves selecting a few individuals or perceptions from a larger population to assemble data and make determinations about the entire population. When it is troublesome or difficult to get data from each individual in the population, this is a helpful and successful strategy for gathering social occasion information.

- **Different sampling methods**

1) **Random Examining:** As the name suggests, information is accumulated randomly while utilising this testing procedure. This indicates that each article in the universe is equally likely to be picked for additional review.

2) **Purposive or Conscious Examining:** Purposive or purposeful testing is the cycle wherein the agent himself

picks the example that best addresses the universe from his viewpoint

3) **Stratified or Blended Inspecting:** Delineated or blended examining is a testing procedure that functions admirably when there are many gatherings inside the populace, each with interesting elements, and a request should be finished on them.

4) **Systematic Examining:** The Efficient Testing Strategy for information assortment requires the organised course of action of different populace units in mathematical, sequential, and geographic requests.

5) Quota Testing: To gather information utilising the Quantity Inspecting Strategy, the entire population is divided into a few classes or groups. Based on the predefined population's many ascribes, it is done. The specialist fixes a couple of rates for the different groupings with different population-wide highlights.

6) **Convenience Testing:** Comfort examining, as the name suggests, is an information assortment procedure wherein the examiner picks the items from the populace that are generally helpful for him.

- **Knowledge Check -1**

  Fill in the Blanks

  1. The act of selecting a subset of individuals or items from a broader population in order to draw inferences about the population as a whole is known as _____ in statistics. (Sampling)
  2. To gather data using the _____ Sampling Method, the total population is split up into several classes or groups. (Quota)

3. _____ sampling is the process wherein the investigator himself chooses the sample that best represents the universe from his perspective. (Purposive)
4. _____ sampling technique covers a wide range of demographic characteristics since it selects groups within a population with varying characteristics. (Systematic)

- **Outcome-Based Activity 1**

You want to launch a new dessert as the head chef at a restaurant. You must compile information about your customers' preferred flavours in order to determine their preferences. Explain the process you would use to choose a representative sample of your clientele, taking into account various sampling techniques in order to guarantee accurate findings. Make sure you provide evidence for your decision by citing the context and the data you hope to get.

9.3 Types of Sampling

- **Two main types of Sampling**

One of these two fundamental classes applies to all inspecting strategies:

1. Probability examining: This method permits specialists to decide the probability that any person in the population will be picked for the review. These examinations provide a more precise and careful numerical examination.
2. Nonprobability examining: In nonprobability testing, scientists can't decide every individual from the populace's likelihood of being remembered for the review. These examples are ordinarily less agent of the entire populace and less exact.

- **Five types of Probability Sampling**

The five likelihood inspecting methods that scientists utilise are as follows:

A) Basic irregular testing

SRS, or basic arbitrary testing, is the cycle wherein each example member has an equivalent possibility of being chosen for the examination. Contemplate utilising a lottery framework. You can pool every single possible responder and pick members aimlessly or indiscriminately. There is an equivalent opportunity for every person in the pool that you will like them. Specialists can likewise utilise PC programs that produce irregular numbers from an assortment.

B) Using delineated testing

A variation on irregular examining hit separated inspecting involves separating the population into a few groups, or layers. The objective of this procedure is to increase the population's test portrayal. Various groups might be remembered for a solitary exploration.

C) Methodical arbitrary inspecting

Scientists utilise efficient inspecting when they select a specific section of individuals as study members in light of a rundown. As an outline, you could make a rundown of 250 individuals in the populace and assign each fifth one as an examination member.

D) Group choice

The method of gathering a populace into bunches or groupings is known as group examination. Groups are often connected with particular geographic districts.

E) Multiple-stage testing

When you utilise a few example methods at different periods of a similar report, it's known as multistage inspecting. This approach is helpful when managing large populations.

- **Types of non-probability Sampling**

Four examples of nonprobability testing are given below:

A. Easy-to-utilize examining

Specialists use an irregular example of individuals for their tests in this examination. A scientist may, for example, take an example of people passing by on the road.

B. Sampling portion-wise

Utilising quantity testing, scientists collect an example as per foreordained qualities. For example, the specialist might gather a partner of people who are 65 years old or more seasoned.

C. Inspecting with judgment

The specialist has unlimited authority over the exploration members they pick while utilising critical testing.

D. Snowball inspecting

At the point when a review requires inspecting gatherings that are more difficult to assemble, scientists use snowball testing.

9.4 Sample Size Determination

- **Meaning of sample size determination**

Deciding the fitting number of perceptions or people to remember for an example from a greater gathering is the most common way of determining example size. Deciding the example size ought to intend to give an example that is both sufficiently enormous to yield measurably huge discoveries and exact populace boundary assessments while being sensible and prudent.

9.5 Sampling in Business Research

In the vast universe of business research, where data is frequently plentiful, testing assumes a basic part. It resembles taking a small, painstakingly picked cut of a monster cake to figure out the kind of the whole. Inspecting permits specialists to assemble information from a reasonable subset of a bigger population while planning to precisely address the whole gathering.

This is the way sampling works in business research:

1. Characterising the Population: The initial step is recognising the whole gathering of interest, called the populace.
2. Test Selection: The following stage is picking the actual example. Different testing strategies exist, each with its assets and shortcomings.
3. Test Size: The size of the example is essential. A bigger example generally prompts more exact outcomes; however, it likewise expands the expense and time expected for information assortment.

- **Knowledge Check -2**

  State True or False

  1. Determining the appropriate number of observations or individuals to include in a sample from a bigger group is the process of determining sample size. (**True**)
  2. Using snowball sampling, researchers assemble a sample according to predetermined characteristics. (**False**)
  3. When a study requires sampling groups of people who are more challenging to gather, researchers utilise snowball sampling. (**True**)
  4. The process of grouping a population into clusters, or groupings, is known as random sampling. (**False**)

- **Outcome-Based Activity -2**

  An investigator seeks to determine the mean age of moviegoers in a specific city. They want to poll a subset of moviegoers. Describe how the sample size needed for the study would change depending on the estimated confidence level and margin of error.

9.6 Summary

- In statistics, the process of gathering data and analysing it to conduct a population study is known as sampling techniques or sample strategies.
- Any of these sampling techniques may involve focusing in particular on hard-to-reach or challenging demographics.
- Systematic sampling, however, may induce bias if the population list has cyclical patterns.
- Researchers may obtain information from a reasonable number of people by choosing a statistically meaningful sample, greatly speeding up the data-gathering process.
- Gathering data from a whole population can have significant financial ramifications, including labour costs, material costs, and overhead.
- According to this concept, each item has an equal and likely probability of being chosen, meaning that the sample selection process is random.

9.7 Keywords

- **Sampling** – The act of selecting a subset of individuals or items from a broader population in order to draw inferences about the population as a whole is known as sampling in statistics
- **Random Sampling** – Every person in the population has an equal probability of getting chosen using this method.
- **Cluster Sampling** – A resource-efficient technique that works well for widely dispersed populations is cluster sampling.
- **Systematic Sampling** – Selecting samples at regular intervals—for example, every tenth person on a list—is known as systematic sampling.

- **Principle of Statistical Regularity** – This concept states that a significant number of randomly picked things from the universe are likely to share the same characteristics as the population as a whole.
- **Principle of "Inertia of Large Numbers** "- According to this idea, conclusions are more likely to be correct the greater the sample size.

9.8 Self-Assessment Questions

1. What are the key sampling techniques?
2. What are the different principles of sampling?
3. What are the five different methods of sampling?
4. What are the 2 main types of Sampling with examples?
5. How to determine sample size?

9.9 References/ Reference Reading

- Chaudhuri, Arijit, and Horst Stenger. *Survey Sampling: Theory and Methods*. 2nd ed., CRC Press, 2023.
- Kumar, Tuli Ram. *Sampling Theory and Testing Hypothesis*. Sage Publications, 2022.
- Mukhopadhyay, Parimal. *Theory and Methods of Survey Sampling*. 3rd ed., PHI Learning, 2023.
- Singh, D. and Chaudhary, F. S. *Theory and Analysis of Sample Survey Designs*. New Age International, 2023.

Unit 10: Sampling and Non-Sampling Errors

Learning Outcomes:

- Students will be able to define the types of sampling errors.
- Students will be able to evaluate how to reduce Sampling errors.
- Students will be able to analyse the differences between sampling errors and non-sampling errors.
- Students will be able to understand the meaning of the Central Limit Theorem.
- Students will be able to remember the techniques to minimise sampling errors.

Structure

10.1 Introduction to Sampling Errors

- **Meaning of sampling errors**

The inborn imperfection in using a small assortment of information (the example) to address a bigger aggregate (the populace) is an examination of mistakes. Picture a bowl loaded with blended nuts. Rather than counting each nut to get the typical cashew rate, you would draw a modest bunch or the example. On the improbable occasion, on the off chance that most of what you wind up getting is peanuts, your gauge of the level of cashews in the example will be off from the genuine extent in the bowl all in all (populace).

Types of Sampling errors

There are various classes of testing mistakes

1) Population explicit blunder

In measurable examinations, populace explicit slip-up, some of the time alluded to as inclusion blunder, happens when there is essential confusion of the objective populace. Consider a specialist who needs to figure out how the public feels about a proposed efficient power energy regulation. They make a telephone overview; however, they limit the example casing to landlines solely, leaving out the rising number of individuals who utilize cell phones.

2) Selection mistake

One sort of inspecting botch is a choice blunder, which compromises the representativeness of your example and, accordingly, the pertinence of your outcomes to the planned populace. It seems that the populace's subgroups are reliably given inclination over others all through the choice interaction.

3) Sample casing mistake

Test outline blunder happens when the reason for picking an example is defective. Consider the example outline as a net that is utilised to gather fish (the populace) from a huge sea. The catch (test) won't accurately mirror the total populace on the off chance that the net is mistakenly made with openings in it or, on the other hand, assuming it is tossed into some unacceptable region of the sea (wrong subpopulation).

4) Non-reaction blunder

One significant overview stress is a non-reaction blunder, which happens when information is missing because a few individuals from the chosen test don't reply. If the qualities of people who don't answer contrast efficiently with those of the individuals who do, then this absence of information brings a predisposition to the discoveries.

There are two sorts of non-reaction mistakes: incomplete and all-out.

- **Examples of sampling errors**

Suppose that XYZ Organisation offers membership-based help that empowers clients to watch films and other content on the web for a set monthly cost.

The organisation is hoping to survey homes that pay for a current video web-based feature and watch about 10 hours of content online every week. XYZ is attempting to figure out to what extent individuals may be keen on a membership administration that costs less. A few sorts of test botches could occur if XYZ doesn't give the examining method enough thought.

Mistakes in populace particular would emerge assuming XYZ Organisation knows nothing about the exact client sorts that should be addressed in the example. For example, on the off chance that XYZ produces a populace of people between the ages of 15 and 25, a huge piece of those clients might not have regular positions, so they don't choose to buy a video web-based feature. Be that as it may, assuming XYZ collected an example of working individuals who are leaders with regard to buys, this arrangement of clients probably won't watch 10 hours of video programming each week.

10.2 Non-Sampling Errors

- **Meaning of non-sampling errors**

Quite possibly, one of the most well-known issues in measurable examination is non-testing botches. Non-testing blunders result from various slip-ups made during the gathering, handling, or examination of information rather than inspection mistakes brought about by the innate furthest reaches of utilising an example to address a more extensive populace. These mix-ups can slant the information and give untrustworthy or one-sided discoveries.

- **Types of non-sampling errors**

No-examining mistakes can cause both arbitrary and precise slip-ups. The main issue is methodical slip-ups. Since arbitrary slip-ups in enormous examples like the ACS will quite often counteract at higher geographic levels, they are less critical.

1) Coverage mistake

Covert age alludes to a lodging unit or individual having no chance of determination in the example. However, over-inclusion alludes to a lodging unit or individual getting a few opportunities of choice in the example or being remembered for the instance when they shouldn't have been.

2) Unit non-reaction

The failure to get the absolute minimum of data from a lodging unit or an occupant in bunch quarters for it to be considered a completed meeting is known as unit nonresponse. When a unit or individual is nonresponsive, it shows that no study information is accessible for that unit or person.

3) Item Non-Reaction

A responder is said to have given a nonresponse if they either didn't address an important inquiry or assumed their reaction went against other data. Thing nonresponse happens when a piece of the study is finished and returned, while replies to different things stay unanswered.

4) Response mistake

Incorrect information announced or recorded is known as a reaction mistake. Reaction mistakes could have been caused by the questioner, the poll, the respondent, or the study strategy itself.

5) Processing Mistake

Blunders could happen when the last information records are being prepared. For example, slip-ups could occur if data from a poll is placed inaccurately or not entirely.

- **Knowledge Check 1**

  Fill in the Blanks

  1. The inherent flaw in utilising a small collection of data (the sample) to represent a larger total (the population) is _____ error. (Sampling)
  2. A _____ error might happen if the possible responder was not contacted or if they declined to answer. (non-response)
  3. When the incorrect subpopulation is chosen to create a sample, _____ error happens, resulting in a sample that noticeably misrepresents the whole population. (Sample frame
  4. _____ nonresponse has a direct impact on the quality of the data; hence, measuring it is crucial. (Unit)

- **Outcome-Based Activity 1**

  A researcher is investigating college students' sleeping patterns. They intend to carry out an online survey. Name two possible non-sampling flaws that can jeopardise the study's accuracy and describe how they might affect the findings.

10.3 Central Limit Theorem

- **Meaning of Central Limit Theorem**

The inspecting conveyance — that is, the likelihood dispersion of a measurement for an enormous number of tests drawn from a populace — is the establishment after that,

as far as possible hypothesis is based. You can have a superior comprehension of test circulations by picturing an investigation:

- **Formula for central limit theorem**

Thankfully, it is possible to determine the sampling distribution's form without actually sampling a population several times. The population's properties dictate the parameters of the sampling distribution of the mean:

A) The population mean is equal to the mean of the sample distribution.

B) The population standard deviation divided by the square root of the sample size yields the standard deviation of the sampling distribution.

With the following notation, we can explain the sampling distribution of the mean:

$$\bar{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$$

Where:

- $\bar{X}$ is the sampling distribution of the sample means
- ~ means "follows the distribution"
- N is the normal distribution
- μ is the mean of the population
- σ is the standard deviation of the population
- n is the sample size

- **Conditions for Central Limit Theorem**

As indicated by as far as possible hypothesis, given the accompanying conditions, the examining circulation of the mean will constantly look like a typical conveyance:

A) There is an adequate amount of information in the example. Commonly, this prerequisite is fulfilled when the example size is $n \geq 30$.

B) The tests are irregular factors that are indistinguishably conveyed and autonomous (i.i.d.). At the point when the example is irregular, this prerequisite is frequently fulfilled.

C) The change of the populace's conveyance is restricted. The Cauchy dissemination is one illustration of a conveyance with a boundless difference to which, as far as possible, the hypothesis doesn't have any significant bearing. The change of most appropriations is restricted.

10.4 Techniques to minimise sampling errors

Testing blunders, the unwanted visitors at the information investigation party can essentially slant the effects of business research. Fortunately, there are strategies

specialists can utilize to limit these excluded mistakes and guarantee their discoveries precisely mirror the larger population.

One key system is to guarantee a sufficient test size. While a greater example isn't generally an assurance of exactness, it improves the probability of catching the genuine variety of the populace.

Another method is to focus on random sampling. This technique gives each individual from the populace an equivalent possibility of being chosen, lessening the gamble of predisposition sneaking in.

Separated sampling is one more weapon in the battle against testing mistakes. Here, the populace is separated into subgroups (layers) in light of pertinent attributes like age, pay, or area.

Limiting non-reaction bias is likewise significant. This happens when a critical part of the picked test doesn't answer overviews, interviews, or different information assortment strategies.

Finally, pilot testing the chosen examining technique can be profoundly useful. By leading a limited-scale preliminary attempt with the chosen examining approach, specialists can identify any possible issues, such as indistinct study questions or trouble arriving at explicit socioeconomics.

- **Knowledge Check -2**

  State True or False

  1. The sampling distribution will resemble a normal distribution more closely with the greater sample size. (True)
  2. When sampling without replacement, the sample size should not exceed 20% of the population. (False)
  3. The number of observations taken from the population for each sample is known as the sample size (n). (True)
  4. The sample size serves as a gauge for the distribution’s variability, or how broad or narrow it is. (False)

- **Outcome-Based Activity -2**

  A researcher would like to know the average salary of teachers in a sizable school district. They want to poll a selection of instructors at random. Describe how this

situation relates to the Central Limit Theorem. Would the study's required sample size be affected by the theorem?

10.5 Summary

- The sampling error increases with sample dissimilarity from the population, which may lead to possibly false conclusions about the population as a whole.
- An analysis known as sampling is carried out by picking a few observations from a larger population.
- Sample frame error, non-response error, selection error, and population-specific error are the four broad categories into which sampling mistakes fall.
- A typical illustration would be a survey that depends exclusively on a tiny percentage of respondents who reply right away.
- Targeting your current clients (a convenience sample) solely while researching social media habits, for example, may skew results in favour of frequent users.
- A non-response error might happen if the possible responder was not contacted or if they declined to answer.

10.6 Keywords

- **Sampling errors** – The inherent flaw in utilising a small collection of data (the sample) to represent a larger total (the population) is sampling error.
- **Coverage error**—In statistical investigations, a population-specific mistake, sometimes referred to as coverage error, occurs when the target population is misinterpreted.
- **Non-Response Error** – One major survey worry is a non-response error, which occurs when data is absent because some members of the selected sample do not answer.
- **Unit Non-Response**—The inability to obtain the bare minimum of information from a housing unit or a resident in group quarters for an interview to be deemed finished is known as unit non-response.
- **Central Limit Theorem** – The sampling distribution—that is, the probability distribution of a statistic for a large number of samples drawn from a population—is the foundation upon which the central limit theorem is based.

10.7 Self-Assessment Questions

1. What are the different types of Sampling errors?
2. What are the steps to reduce Sampling errors?
3. What are the types of non-sampling errors?
4. What are the major differences between sampling errors and non-sampling errors?
5. What are the conditions and assumptions of the Central Limit Theorem?

10.8 References/ Reference Reading

- Cochran, William G. *Sampling Techniques*. 3rd ed., John Wiley & Sons, 2007.
- Lohr, Sharon L. *Sampling: Design and Analysis*. 2nd ed., Cengage Learning, 2010.
- Thompson, Steven K. *Sampling*. 3rd ed., John Wiley & Sons, 2012.
- Murthy, M.N. *Sampling Theory and Methods*. 2nd ed., Statistical Publishing Society, 2019.
- Chaudhuri, Arijit, and Horst Stenger. *Survey Sampling: Theory and Methods*. 2nd ed., CRC Press, 2005.

Unit 11: Tests of Hypothesis

Learning Outcomes

- Students will be able to define hypothesis testing.
- Students will be able to evaluate possible difficulties in the formulation of a good hypothesis.
- Students will be able to analyse different types of hypothesis testing.
- Students will be able to understand the statistical significance of the p-value.
- Students will be able to remember applications of hypothesis testing in business research.

Structure

11.1 Introduction to hypothesis testing
- Meaning of hypothesis testing
- Importance of hypothesis formulation

11.2 Formulating Hypotheses
- Meaning and characteristics of hypothesis
- Formulation of hypotheses
- Knowledge Check 1
- Outcome-Based Activity 1

11.3 Types of hypothesis tests

11.4 p-value and statistical significance

- Meaning of p-value
- Statistical significance of p-value
- Knowledge Check Activity 2
- Outcome-Based Activity 2

11.5 Summary

11.6 Keywords

11.7 Self-Assessment questions

11.8 References/ Reference Reading

11.1 Introduction to Hypothesis Testing

- **Meaning of hypothesis testing**

A technique for drawing statistical conclusions about the population data is hypothesis testing. It is a tool for analysis that verifies hypotheses and assesses the likelihood of events falling within a predetermined range of accuracy. Testing hypotheses offers a means of confirming the validity of experiment findings. Prior to conducting the hypothesis test, a null hypothesis and an alternative hypothesis are established. This facilitates concluding the demographic sample that was taken. A statistical technique called hypothesis testing is employed to determine the significance of experiment outcomes.

- **Importance of hypothesis formulation**

A hypothesis is the fundamental purpose of scientific study. In the event that a concise, unambiguous scientific theory has been developed, the investigator will have little trouble moving forward with the investigation. This is how its usefulness or significance for study may be examined. By identifying the precise phenomena being studied, a well-defined hypothesis gives attention.

Goode and Hatt ('without' hypothesis formulation) state that research is aimless and consists of haphazard empirical meandering. The findings cannot be analysed as definitive facts. The creation of hypotheses connects theory with research, resulting in the discovery of new information. They contend that without structure, the study devolves into arbitrary empirical meandering—a meaningless stroll through data. This directionless produces results that are difficult to evaluate as true categorically.

11.2 Formulating Hypotheses

- **Meaning and characteristics of hypothesis**

After a research topic is formulated, the creation of a hypothesis is a crucial step in the research process. As you are aware, the first step in every scientific investigation is to identify a problem that can be solved. Once the issue has been identified, the researcher presents a potential solution in the form of a tested hypothesis. A hypothesis is sometimes seen as a tentative and verifiable assertion of a possible connection between two or more variables or events that are being studied.

"A testable statement of a potential relationship between two or more variables, i.e., advance as a potential solution to the problem," is what Mcguigan (1990) defined as a hypothesis. "A hypothesis is a conjectural statement of the relation between two or more variables," according to Kerlinger (1973). A hypothesis must be formulated so that it can be empirically tested in order for it to be valuable in any investigation. A few of these attributes are listed as follows:

1) A hypothesis should have a clear conceptual framework
2) It should be testable
3) It should have relevance to the body of current knowledge and its implications
4) It should be logically cohesive and comprehensive
5) It should be verifiable
6) It should be operationalised.

- **Formulation of Hypotheses**

The route of scientific inquiry is systematic and evidence-based. Researchers start by making observations and looking for patterns or phenomena that catch their attention. This observation then sets the stage for the creation of hypotheses, an important process that establishes a provisional explanation for the phenomena being seen. The investigation is guided by this hypothesis, which is a precise and verifiable assertion regarding the relationship between the variables.

- **Knowledge Check 1**

  Fill in the Blanks

1. A technique for drawing statistical conclusions about the population data is _____ testing. (hypothesis)

2. A hypothesis is a conjectural statement of the relation between two or more variables," according to _____ (1973). (Kerlinger)
3. Russell and Reichenback (_____) recommended that the hypotheses be presented logically with regard to the broad implications. (1947)
4. The route of _____ inquiry is systematic and evidence-based. (Scientific)

- **Outcome-Based Activity 1**

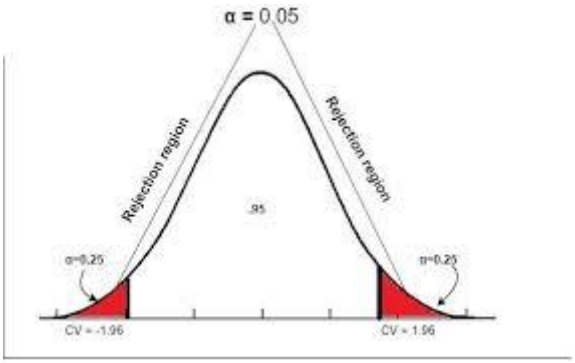
  An investigator wants to look into how well a novel fertiliser promotes plant development. Create a concise and unambiguous hypothesis that may be investigated in an experiment. Include the factors as well as the anticipated relationship between them.

11.3 Types of hypothesis tests

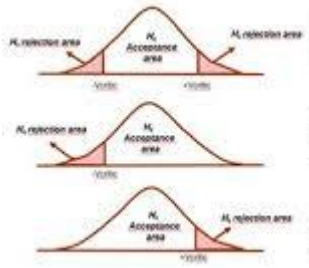
There are 2 main ways to categorize hypothesis tests: -

1) By Direction of the Alternative Hypothesis- This spotlight on what the elective speculation (Ha) is attempting to demonstrate. There are 3 main types: -

a. Two-tailored tests- This is the broadest sort of speculation test. The elective speculation (Ha) expresses that the populace boundary of interest is not the same as a particular worth proposed by the invalid theory. Here, the impact could be either bigger or minuscule.



b. Right-tailed test—This kind of test is utilised when the specialist expects a beneficial outcome, meaning the population boundary is more noteworthy than the worth proposed by the invalid speculation.



H0: sample mean = population mean
Ha: sample mean <> population mean

H0: sample mean >= population mean
Ha: sample mean < population mean

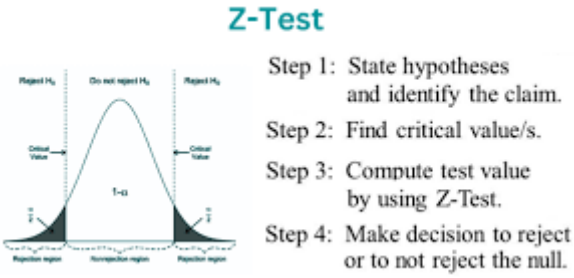
H0: sample mean <= population mean
Ha: sample mean > population mean

c. Left-tailed test- As opposed to the right-followed test, this situation expects an adverse consequence, meaning the populace boundary is not exactly the worth proposed by the invalid speculation. Consider another drug intended to bring down the pulse. The invalid speculation may be that the medicine significantly affects circulatory strain. The left-followed elective speculation, nonetheless, recommends that the medicine will really diminish circulatory strain contrasted with the pattern level.

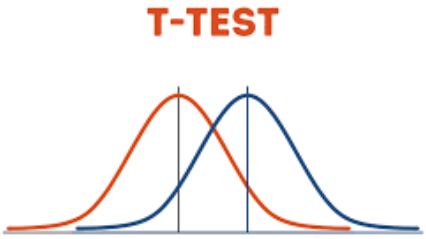
2) By Type of Statistical Test- This alludes to the particular factual technique used to investigate the information. Normal models include:

a. Z- test- This workhorse of speculation testing is utilised while contrasting the method for two gatherings, expecting the information to be regularly circulated and the populace standard deviation to be known. It assists us with deciding whether the noticed distinction between the two gatherings is logical because of irregular possibility or mirrors a genuine basic contrast in the populace. The z-measurement, a normalised score, is determined in light of the example means and standard deviation.

b.



c. T-test—Like the z-test, the t-test is used to examine implies; however, it proves useful when the population standard deviation is obscure. In such cases, the example standard deviation is used to gauge the population's esteem, making the t-circulation more suitable than the ordinary conveyance. The t-test understands a comparative rationale to the z-test, computing a t-measurement and contrasting it with a t-dissemination table with the fitting levels of opportunity (in light of test size) to decide the p-esteem.



d. Chi-square test—This test is utilised to survey the connection between two unmitigated factors. It assists us with deciding whether the noticed circulation of frequencies across classes is possible because of the possibility of mirrors a veritable relationship between the factors. The chi-square measurement is determined in light of the noticed and anticipated frequencies for every classification mix.

Chi-square test

$$\chi^2 = \frac{\sigma s^2}{\sigma p^2}(n-1)$$

$$\chi^2 = \sum \frac{(Oij - Eij)^2}{Eij}$$

11.4 p-value and statistical significance

- **Meaning of p-value**

In the realm of measurable speculation testing, the p-value holds gigantic significance. It resembles a criminal investigator's likelihood scale, assisting specialists with measuring the probability of a specific result emerging by irregular possibility. We should dig further into the importance of p-worth and how it's deciphered. Envision you're directing a review and suspect (invalid theory) that there's no genuine distinction between two gatherings. The p-esteem lets you know the likelihood of noticing your review's outcomes (or significantly more outrageous outcomes), assuming that the underlying hunch was valid. It resembles working out the chances of haphazardly picking a dark marble from a pack; just here, the sack contains a blend of proof supporting or discrediting your hunch.

Here's the reason:

a. Test size matters: A bigger example size can prompt lower p-values for trifling contrasts. In this way, the setting of your review and the extent of the impact are pivotal.
b. Irregular chance: Even with a low p-esteem, there's dependably a little chance the outcome is because of possibility. Replication of studies sets discoveries.

- **Statistical significance of p-value**

In the domain of factual speculation testing, the p-value arises as a basic player. It connotes the likelihood of getting results as outrageous or more limit than what was noticed, expecting the invalid speculation (the default supposition of no impact) to be valid. How about we dive further into the universe of p-values and their importance in deciphering research discoveries?

The p-value doesn't straightforwardly let us know whether a finding is valid or bogus, but rather the strength of proof against the invalid speculation. A lower p-esteem demonstrates a more grounded case for dismissing the invalid speculation. Here's the reason deciphering them cautiously is fundamental:

1. Edge Dependence: The importance level (frequently 0.05) is a pre-decided benchmark. A marginally higher p-esteem (e.g., 0.06) doesn't be guaranteed to compare to an unimportant outcome. The significance of a finding could rely upon the exploration setting, regardless of whether it falls simply over the conventional edge.
2. Test Size Dependence: The p-worth can be delicate to test size. With bigger examples, even little impacts can yield genuinely critical outcomes. On the other hand, more modest investigations could miss genuine implications because of a higher p-esteem.

11.5 Applications of hypothesis testing in business research

Hypothesis testing is an incredible asset utilised widely in business exploration to pursue information-driven choices and gain important bits of knowledge. It permits specialists to move past essentially portraying information to assess the legitimacy of cases and connections between factors. Here is a more critical glance at how theory testing is applied in different business research situations:

1. Market Research: Imagine an organisation that needs to test the viability of another promotion effort. They can speculate that the new mission will prompt higher brand mindfulness compared with the last mission.

2. Item Development: Organisations frequently improve and foster new items. Speculation testing assumes an urgent part in assessing new item elements or functionalities.
3. Human Resources: Advancing human asset rehearses is fundamental for business achievement. Speculation testing can be utilised to survey the viability of preparing programs. Scientists could speculate that another preparation program will prompt better worker execution.

- **Knowledge Check Activity 2**

  State True or False

  1. Hypothesis testing is an incredible asset utilised widely in business exploration to pursue information-driven choices and gain important bits of knowledge. (True)
  2. A genuinely huge outcome with more impact size could hold less down-to-earth importance. (False)
  3. A low p-value (ordinarily beneath 0.05) demonstrates it's implausible to come by such an outrageous outcome in the event that there's genuinely no distinction between the gatherings. (True)
  4. The t-measurement, a normalised score, is determined in light of the example means and standard deviation. (False)

- **Outcome-Based Activity 2**

  A showcasing group runs an online entertainment mission to advance another sound café. They gather information on site visits during the mission and play out a speculation test. The p-esteem from the test is 0.02. Based on this p-value, could you at any point make sense of the event that the showcasing effort altogether affected site traffic? Legitimise your response by making sense of the idea of p-esteem and measurable significance.

11.6 Summary

- Prior to conducting the hypothesis test, a null hypothesis and an alternative hypothesis are established.

- Setting up a test to see if a new medication treats a condition more effectively is an example of hypothesis testing.
- It's a systematic approach of using information gathered from a smaller sample to make trustworthy inferences about a wider population.
- Since it is impossible to accept a hypothesis with 100% precision, we often choose a significance threshold of 5%.
- The likelihood of discovering the observed/extreme outcomes when the null hypothesis (H0) of a topic provided for investigation is correct is known as the P value or computed probability.
- When making inferences about a population from a sample of data, researchers may commit Type I and Type II mistakes in hypothesis testing.
- A hypothesis must be formulated so that it can be empirically tested in order for it to be valuable in any investigation.
- The elective speculation (Ha) expresses that the populace boundary of interest is not the same as a particular worth proposed by the invalid theory.

11.7 Keywords

- **Hypothesis testing**- It is a tool for analysis that verifies hypotheses and assesses the likelihood of events falling within a predetermined range of accuracy.
- **Test Statistic**- In a hypothesis test, the test statistic is a number that is computed from sample data and is used to decide whether to reject the null hypothesis.
- **Two tailored tests**—The elective speculation (Ha) expresses that the populace boundary of interest is not the same as the particular worth proposed by the invalid theory.
- **Right-tailed test**—This kind of test is utilised when the specialist expects a beneficial outcome, meaning the population boundary is more noteworthy than the worth proposed by the invalid speculation.
- **Left-tailed test—**As opposed to the right-followed test, this situation expects an adverse consequence, meaning the populace boundary is not exactly the worth proposed by the invalid speculation.
- **Chi-Square test**- This test is utilised to survey the connection between two unmitigated factors.

11.8 Self-assessment questions

1. What are the type 1 and type 2 errors in hypothesis testing?
2. What are the possible difficulties in the formulation of a good hypothesis?
3. What are the different types of hypothesis tests?
4. What is the statistical significance of the p-value?
5. What are the different applications of hypothesis testing in business research?

11.9 References/ Reference Reading

- Gupta, S. P., and M. P. Gupta. *Business Statistics*. 18th ed., Sultan Chand & Sons, 2014.
- Keller, Gerald. *Statistics for Management and Economics*. 11th ed., Cengage Learning, 2017.
- Sharma, J. K. *Business Statistics: Problems and Solutions*. 2nd ed., Pearson Education India, 2016.
- Aczel, Amir D., and Jayavel Sounderpandian. *Complete Business Statistics*. 8th ed., McGraw-Hill Education, 2012.
- Arora, P. N., S. Arora, and Amit Arora. *Comprehensive Statistical Methods*. 3rd ed., S. Chand Publishing, 2010.

Unit 12: Index Numbers

Learning outcomes

- Students will be able to define the meaning of index numbers.
- Students will be able to evaluate the methods of constructing index numbers.
- Students will be able to analyse the problems in the construction of index numbers.
- Students will be able to understand the major limitations of index numbers.
- Students will be able to remember the key differences between the consumer price index and the wholesale price index.

Structure

12.1 Characteristics and Utility of Index Numbers

- **Meaning of Index Numbers**

The index number was first developed by an Indian Analyst, Carli, in 1764. It was utilised interestingly to analyse the costs of 1750 with those of 1500. An index number is a factual device for estimating changes in the greatness of a gathering of related factors.

"index Number shows by its variety the progressions in a size which isn't vulnerable both of precise estimation in itself or of direct valuation by and by."- Edgeworth.

"Index numbers are gadgets for estimating contrasts in the size of a gathering of related factors."- Croxton and Cowden.

"An index number is a factual measure intended to show changes in a factor or a gathering of related factors concerning time, geological area or different qualities." – Spiegel.

- **Characteristics of Index Numbers**

The different characteristics of Index Numbers are as per the following:

1. Particular Midpoints:

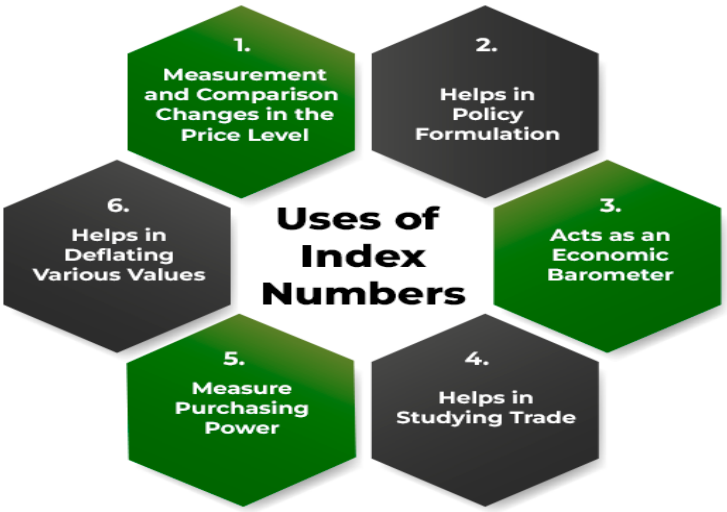
While mean, median, and mode offer important outlines for single datasets; their constraints become obvious while examining connections between information with blended kinds or units. Envision concentrating on plant development, with information on level (centimetres) and leaf count. The typical level (20 cm) and most successive leaf count (10) don't uncover, assuming taller plants have more leaves. These midpoints expect a reliable information type, which isn't true here.

2. Communicated in Rates:

The greatness of a gathering's progressions is communicated as far as rates, which are free of the estimation units. It helps in sorting out how at least two file numbers are analysed in different conditions. However, the % sign is rarely utilised.

- **Utility of Index Numbers**

Basically, in every aspect of monetary action, changes are estimated utilising list numbers. It helps in keeping changes in yield, pay, work, business exercises, efficiency, and so on. As indicated by M.M. Blair, "File numbers are the signs and guide posts along the business thruway that demonstrate to the financial specialist how he ought to drive or oversee undertakings.



1. Estimation and Examination of Changes in the Cost Level:

It is beyond the realm of possibilities to expect to quantify changes in the value level of two factors in outright terms; in this manner, record numbers give a general measure to the progressions in the greatness of a gathering of factors.

2. Helps in Strategy Definition:

An index number is a significant device for government or non-government associations in the accompanying ways: In a strategy plan, there is a requirement for a base or pattern. With the index numbers, the patterns of various peculiarities can be contemplated.

12.2 Methods of Constructing Index Numbers

There are two methods through which we can develop Index Numbers: viz., Straightforward or Unweighted and Weighted Index Numbers.

1. Unweighted or Straightforward Index Numbers

An Index Number in which everything should have some weight as no weight is explicitly appointed to anything is known as an Unweighted List Number. It is otherwise called a Straightforward List Number and can be built with the assistance of two procedures, viz., Simple Aggregative Method and Simple Average of Price Relatives Method.

a) Simple Aggregative Method:

A strategy where the total costs of the multitude of chosen wares in the ongoing year are communicated as the total cost of the relative multitude of products in the base year is known as the Simple Aggregative Method. It is the least complex technique to build file numbers. The equation for developing the Index number is:

$$P_{01} = \frac{\sum P_1}{\sum P_0} \times 100$$

Example—From the accompanying information, build Index Numbers for 2018-2019, requiring 2010-2011 as the base year, using the Simple Aggregate Method.

Commodity	**Price in 2010-2011(Rs/l)**	**Price in 2018-19(Rs/l)**
Milk	50	70
Juice	30	40
Shake	70	90
Smoothie	100	50

Construction Of Price Index

Commodity	**Price in 2010-2011**	**Price in 2018-19**
Milk	50	70
Juice	30	40
Shake	70	90
Smoothie	100	150
	Sum Total= 250	Sum Total= 350

Price Index for the years 2018-2019 with the years 2010-2011 as the base: 140. The worth of the Price Index (140) uncovers that in the year 2018-2019, there is a net increment of 40% in the costs of given wares when contrasted with the year 2010-2011.

b) Simple Average of Price Relatives Method

This index number is an improvement over the past strategy, as it isn't impacted by the unit in which the costs of the wares are cited. The equation for developing index number is:

$$P_{01} = \frac{\sum\left(\frac{P_1}{P_0} \times 100\right)}{N}$$

2) Weighted Index Numbers

In other index numbers, equivalent significance is given to everything. Notwithstanding, under the Weighted Index Number, reasonable loads are given to every ware or thing expressly. These loads show the general importance of the wares or things remembered for the assurance of the record.

A) Weighted Aggregative Method

This method includes relegating loads to various things and getting a weighted total of the costs as opposed to tracking down a straightforward total of costs. A few significant strategies for developing the Weighted Aggregative Method are as follows:

- Laspeyres's Technique- The method of calculating Weighted Index Numbers under which the base year quantities are used as weights of different items is known as Laspeyres's Method. The formula for Laspeyres's Price Index is:

$$p_{01} = \frac{\sum p_1 q_0}{\sum p_0 q_0} \times 100$$

- **Paasche's Technique**- The technique for working out Weighted Index Numbers under which the ongoing year's amounts are utilised as loads of various things is known as Paasche's Strategy. The formula for Paasche's Price Index is:

$$p_{01} = \frac{\sum p_1 q_1}{\sum p_0 q_1} \times 100$$

- **Fisher's Optimal Technique**- The strategy for computing Weighted Index Numbers under which the joined procedures of Paasche and Laspeyres are utilised is known as Fisher's Technique. All in all, both the base year and current year's amounts are used as loads. The formula for Fisher's Price Index is:

$$p_{01} = \sqrt{\frac{\Sigma P_1 q_0}{\Sigma P_0 q_0} \times \frac{\Sigma P_1 q_1}{\Sigma P_0 q_1}} \times 100$$

B) Weighted Average of Price Relatives Method

Under this technique, the base year costs of the products are taken as the premise to work out the cost of family members for the ongoing year. The determined value family members are then duplicated by their separate loads of things. The formula for developing index number is:

$$\text{(b)}\ \frac{\sum p_1 (p_0\, q_0)}{\sum (p_0\, q_0)} \times 100$$

$$\text{(c)}\ \frac{\sum p_0 (p_0\, q_0)}{\sum (p_0\, q_0)} \times 100$$

$$\text{(d)}\ \frac{\sum \left(\frac{p_1}{p_0}\right) (p_1\, q_1)}{\sum (p_1\, q_1)} \times 100$$

12.3 Problems in the Construction of Index Numbers

1) Difficulties in the selection of base period

Picking a base year for monetary markers like expansion presents a consistent test. Preferably, the base year ought to address a time of typical monetary movement, giving a steady reference highlight correlation. Be that as it may, characterising "typical" is intrinsically emotional. A year that shows up financially common at the time could later be viewed as uncommon because of unexpected occasions or long-haul patterns

2) Difficulties in the selection of commodities

Picking the right crate of merchandise for an index number, similar to those following expansion, is a consistent battle. It's not just about tossing in anything that's discounted. These wares need to mirror the average acquisition of most buyers.

- **Knowledge Check 1**

  Fill in the Blanks

  1. A _____is an instrument that is utilised to quantify environmental tension. (Gauge)
  2. A _____number is a factual device for estimating changes in the greatness of a gathering of related factors. We are a piece of a quick-moving economy. (Index)
  3. The index number was first developed by an Indian Analyst, Carli, in _____. (1764)

4. The method of calculating Weighted Index Numbers under which the base year quantities are used as weights of different items is known as _____ Method. (Laspeyres's)

- **Outcome-Based Activity 1**

  You are a monetary investigator entrusted with clearing up the constraints of index numbers for a client. The client runs a little pastry kitchen and is worried about rising flour costs. Make sense of how the decision of the base year and delegate things can affect the exactness of an index number in mirroring the particular cost changes pertinent to their business.

12.4 Limitations of Index Numbers

1. Selection of base year- Choosing a base year for an index number, similar to those following expansion, is a consistent test. In a perfect world, the base year ought to address a time of ordinary financial movement, giving a steady reference highlight examination. Be that as it may, characterising "ordinary" is innately emotional. A year that shows up financially regular at the time could later be viewed as strange because of unexpected occasions or long-haul patterns.

2. Representative Items- Choosing a delegated bushel of merchandise for an index number is a consistent fight against changing purchaser inclinations. In a perfect world, the bushel ought to mirror the common acquisition of the larger part. In any case, what individuals purchase can change fundamentally over the long run. New items hit the market, tastes shift, and, surprisingly, laid-out products could go through quality changes. Envision depends on a record that actually incorporates buggy whips! To keep up with the importance, the bushel of merchandise should be refreshed intermittently. This guarantees it stays lined up with current utilisation designs.

3. Data Collection- Gathering exact and adequate cost information for index numbers presents a critical strategic test. Preferably, we'd need reliable evaluating data from a similar hotspot for every benefit.

12.5 Consumer Price Index and Wholesale Price Index

- **Meaning of Consumer Price Index**

The Consumer Price Index, frequently condensed as CPI, is a significant monetary pointer that mirrors the changing cost for many everyday items for shoppers. Envision it as a goliath shopping crate loaded up with a painstakingly picked choice of labour and products that a normal shopper could purchase consistently. By following the value changes of this bushel after some time, the CPI basically lets us know the amount more (or less) we'd have to spend today to keep up with a similar way of life we delighted in a past period.

The CPI is a fundamental instrument in light of multiple factors:

1. Expansion Check: It's the essential measurement for estimating expansion. By checking CPI changes, policymakers can survey the general strength of the economy and make important moves, such as changing loan costs, to control expansion.
2. Typical cost for many everyday items Changes: The CPI is utilised to change wages, annuities, and government-managed retirement advantages to stay up with expansion and guarantee that their buying power doesn't lessen over the long run.

- **Meaning of Wholesale Price Index**

The Wholesale Price Index (WPI) is a key financial pointer that tracks cost variances for products exchanged massively between organisations, principally makers and wholesalers. Dissimilar to the Consumer Price Index (CPI), which centres around the expense of completed products for purchasers, the WPI gives a window into cost changes at a prior stage in the store network.

The WPI serves several crucial purposes:

1. Early Expansion Signal: Since the WPI tracks costs at a prior stage in the store network, it can go about as an early advance notice sign for possible expansion. Rising discount costs frequently portend expansions in purchaser costs down the line.
2. Business Arranging: Organisations, particularly makers and wholesalers, depend on the WPI to settle on informed conclusions about creation expenses and estimating procedures. Rising discount costs can provoke changes in cycles or item value to keep up with benefits.

- **Knowledge Check 2**

  State True or False

  1. The Wholesale Price Index (WPI) is a key financial pointer that tracks cost variances for products exchanged massively between organisations, principally makers and wholesalers. (True)
  2. By following the value changes of this bushel after some time, the WPI basically lets us know the amount more (or less) we'd have to spend today to keep up with the similar way of life we delighted in a past period. (False)
  3. To explore this intricacy, index numbers frequently depend on discount costs. (True)
  4. the consumer bears WPI. (False)

- **Outcome-Based Activity 2**

  You are a monetary expert entrusted with making sense of the distinctions between the Consumer Price Index (CPI) and the Wholesale Price Index (WPI) to a client clothing-producing business. Make sense of what the CPI and WPI can mean for the client's business in terms of creation expenses and last selling costs.

12.6 Summary

- An index number is a factual device for estimating changes in the greatness of a gathering of related factors.
- "An index number is a factual measure intended to show changes in a factor or a gathering of related factors concerning time, geological area or different qualities." – Spiegel
- It tracks changes in the worth of factors like the typical cost for most everyday items, the volume of creation in different ventures, the costs of a rundown of characterised wares, etc.
- You could analyse the typical expense of normal things in two urban areas simultaneously, uncovering potential cost contrasts between locales.
- In any case, by zeroing in on the relative changes - the varieties in costs across various areas and over the long run - we can uncover the impact of these concealed elements.

- It helps anticipate future patterns, which is essential for any business or creation movement's future tasks.
- Moreover, in the event of a rise in costs, the course of change is known as collapsing.

12.7 Keywords

- **Index Numbers**—An index number is a factual device for estimating changes in the greatness of a group of related factors.
- **Simple Aggregative Method**—This strategy communicates the total costs of the multitude of chosen wares in the ongoing year as the total cost of the relative multitude of products in the base year.
- **Weighted Aggregative Method**—This method involves relegating loads to various things and calculating a weighted total of the costs rather than tracking down a straightforward total of costs.
- **Paasche's method**—The technique for working out Weighted Index Numbers under which the current year's amounts are utilised for loads of various things is known as Paasche's Strategy.
- **Fisher's Optimal Technique**—Fisher's technique is the strategy for computing Weighted Index Numbers under which the joined procedures of Paasche and Laspeyres are utilised.
- **Weighted Average of Price Relative Method**—Under this technique, the base year costs of the products are used to calculate the cost to family members for the ongoing year.
- **Wholesale Price Index**- The Wholesale Price Index (WPI) is a key financial pointer that tracks cost variances for products exchanged mass between organisations, principally makers and wholesalers

12.8 Self-Assessment Questions

1. What are the major characteristics of index numbers?
2. What are the different methods of constructing index numbers?
3. What are the major problems in the construction of index numbers?
4. What are the several limitations of index numbers?

5. What are the major differences between the consumer price index and the wholesale price index?

12.9 References/ Reference Reading

- Dhingra, I. C., and Vishal Agarwal. *Statistics for Economics*. Sultan Chand & Sons, 2022.
- Gupta, S. P. *Statistical Methods*. Sultan Chand & Sons, 2021.
- Kapoor, V. K., and Rajeev Manga. *Business Statistics*. S. Chand Publishing, 2022.
- Lind, Douglas A., William G. Marchal, and Samuel A. Wathen. *Statistical Techniques in Business and Economics*. McGraw-Hill Education, 2020.
- Spiegel, Murray R., and Larry J. Stephens. *Schaum's Outline of Statistics*. McGraw-Hill Education, 2018.

Unit 13: Non-Parametric Tests

Learning Outcomes

- Students will be able to define non-parametric tests.
- Students will be able to analyse the formula used in the Chi-Square test.
- Students will be able to apply the steps involved in the Mann Whitney U Test.
- Students will be able to understand the assumptions of the Kruskal Wallis Test.
- Students will be able to remember the steps involved in the Wilcoxon signed-rank test.

Structure

13.1 Introduction to Parametric Tests

- **Meaning of Parametric Tests**

Non-parametric tests are numerical techniques utilised in factual speculation testing. They don't make presumptions about the recurrence conveyance of factors that are to be assessed. Non-parametric analysis is utilised when there is slanted information, and it includes procedures that don't rely upon information relating to a specific conveyance.

- **Advantages and Disadvantages of Non-Parametric Tests**

Advantages

- Relaxed Supposition about Information Distribution: Dissimilar to parametric tests, which require the information to follow a particular dissemination (frequently typical), non-parametric tests sparkle when your information doesn't conveniently fit a ringer bend.
- Accommodating More Modest Example Sizes: Parametric tests frequently have a base example size that is necessary for their outcomes to be genuinely substantial. This can be an obstacle in research settings where it is costly or illogical to gather a lot of information.
- Suitable for Ordinal and Ostensible Data: Not all information comes in flawless mathematical qualities. In some cases, we manage rankings (ordinal information) like consumer loyalty evaluations (extremely fulfilled, fulfilled, nonpartisan) or groupings (ostensible information) like favoured dress brands (A, B, C)—parametric tests battle with these information types. Non-parametric tests, in any case, are explicitly intended to deal with them.

Disadvantages

The following are four impediments of non-parametric tests made sense of exhaustively:

- Less Powerful: While non-parametric tests are perfect for information that abuses presumptions of ordinariness, they can be less powerful contrasted with their parametric partners when those suspicions are really met. Parametric tests influence

more data about the information circulation, similar to the mean and standard deviation.

- Limited Information: Parametric tests frequently give more detailed information about the information contrasted with non-parametric tests. This is on the grounds that they gauge explicit boundaries of the hidden conveyance
- Less Efficient: Non-parametric tests can be less efficient than parametric tests while managing ordinary information and bigger examples. Productivity alludes to the capacity of a test to recognise a genuine contrast accurately.

13.2 Chi-Square Test

- **Meaning of Chi-square test**

The Chi-Square test is a measurable system for deciding the contrast between noticed and anticipated information. This test can likewise be utilised to determine if it connects to the straight-out factors in our information. It assists with seeing if a distinction between two clear-cut factors is because of possibility or a connection between them. A chi-square test is a measurable test that is utilised to look at noticed and anticipated results.

It is utilised to work out the distinction between two all-out factors, which are:

1. Because of possibility or
2. In view of the relationship

- **Formula for Chi-Square Test**

$$x_c^2 = \frac{\Sigma\,(O_i - E_i)^2}{E_i}$$

Were

C = Levels of opportunity

O = Noticed Worth

E = Anticipated Worth

The levels of opportunity in a measurable estimation address the number of factors that can shift in a computation. The levels of opportunity can be determined to guarantee that chi-square tests are measurably substantial.

- **Steps to perform Chi-Square Test**

The specific system for playing out a Pearson's chi-square test relies upon which test you're utilising, yet it, for the most part, follows these means:

1. Make a table of the noticed and anticipated frequencies. This can, in some cases, be the most troublesome step since you should cautiously consider which expected values are generally proper for your invalid speculation.
2. Ascertain the chi-square worth from your noticed and expected frequencies utilising the chi-square recipe.
3. Find the basic chi-square worth in a chi-square basic worth table or utilising factual programming.
4. Contrast the chi-square worth with the basic worth to figure out which is bigger.
5. Choose whether to dismiss the invalid speculation. You ought to dismiss the invalid speculation if the chi-square worth is more noteworthy than the basic worth. Assuming you reject the invalid speculation, you can infer that your information is fundamentally not the same as what you anticipated.

13.3 Mann- Whitney U Test

- **Meaning of Mann- Whitney U Test**

The Mann Whitney U test can be considered a strong Nonparametric test and a compelling option in contrast to the autonomous T-test. It very well may be utilised satisfactorily with nonstop and discrete factors and can actually be utilised with little example.

The suppositions of Non-parametric tests are pertinent to the Whitney U test. Allow us now to examine a portion of the suppositions of Mann Whitney U test:

1. The perceptions should be free.
2. The reliant variable that is taken in the exploration needs to show coherence.
3. The test is utilised when the two example subgroups are free, and the information is ordinal. On the off chance that the information is in span or proportion structure, the same is changed to rank request.
4. The example sub bunches should be free and not corresponded.

- **Formula for Mann Whitney U Test**

Female

Number of cases: $n_1 = 5$ Rank sum: $T_1 = 28.5$

$$U_1 = n_1 \cdot n_2 + \frac{n_1 \cdot (n_1 + 1)}{2} - T_1$$
$$= 5 \cdot 6 + \frac{5 \cdot (5 + 1)}{2} - 28.5$$
$$= 16.5$$

Male

Number of cases: $n_2 = 6$ Rank sum: $T_2 = 37.5$

$$U_2 = n_1 \cdot n_2 + \frac{n_2 \cdot (n_2 + 1)}{2} - T_2$$
$$= 5 \cdot 6 + \frac{6 \cdot (6 + 1)}{2} - 37.5$$
$$= 13.5$$

Number of all cases

$$n = n_1 + n_2 = 11$$

U-value

$$U = \min(U_1, U_2) \quad = \min(16{,}5\ , 13{,}5) \ = 13.5$$

Expected value of U

$$\mu_u = \frac{n_1 \cdot n_2}{2} \quad = \frac{6 \cdot 5}{2} = 15$$

Standard error of U

$$\sigma_{U_{corr}} = \sqrt{\frac{n_1 \cdot n_2}{n \cdot (n-1)}} \cdot \sqrt{\frac{n^3 - n}{12} - \sum_{i=1}^{k} \frac{t_i^3 - t_i}{12}}_i$$

$$\sigma_{U_{corr}} = \sqrt{\frac{5 \cdot 6}{11 \cdot (11-1)}} \cdot \sqrt{\frac{11^3 - 11}{12} - 2{,}5} \ = 5.41$$

z-value

$$z = \frac{U - \mu_U}{\sigma_{u_{corr}}} \quad = \frac{13.5 - 15}{5.41} \quad = -0.28$$

There are really two adaptations of the U measurement, U1 and U2, contingent upon which gathering's number of positions you choose to use in the equation. Notwithstanding, paying little heed to which form you pick, the end product (either U1 or U2) will constantly be numerically the same and lead to similar decisions about the factual meaning of the contrast between the gatherings.

- **Knowledge Check 2**

  Fill in the Blanks

  1. _____ tests are the numerical techniques utilised in factual speculation testing, which don't make presumptions about the recurrence conveyance of factors that are to be assessed. (Non-parametric)
  2. The _____test is a non-parametric test, meaning it doesn't depend on suspicions about the fundamental information conveyance (like ordinariness). (Mann Whitney U Test)
  3. The _____ is a measurable system for deciding the contrast between noticed and anticipated information. (Chi-Square test)
  4. Parametric tests frequently give more _____information about the information contrasted with non-parametric tests. (Detailed)

- **Outcome-Based Activity 1**

A promoting director is examining the viability of a virtual entertainment crusade focusing on two different age gatherings (18-25 and 26-34). The mission is planned to increment brand mindfulness. The director gathered information on the number of individuals from each age who saw the promotion, loved the promotion, and shared the advertisement. Might you at any point utilize a Chi-Square test to investigate this information and decide whether there's a measurably huge relationship between age gathering and commitment with the web-based entertainment crusade? Explain your reasoning.

13.4 Kruskal- Wallis Test

- **Meaning of Kruskal- Wallis test**

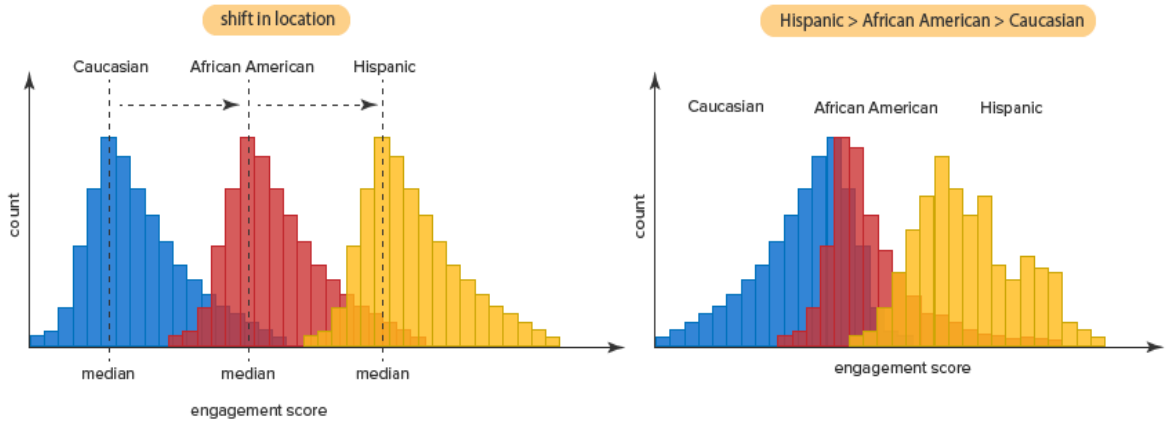
The Kruskal-Wallis H test (at times likewise called the "one-way ANOVA on positions") is a position-based nonparametric test that can be utilised to decide whether there are genuinely massive contrasts between at least two gatherings of a free factor on a consistent or ordinal ward variable. It is viewed as the nonparametric option in comparison to the one-way ANOVA and an expansion of the Mann-Whitney U test to permit the examination of multiple autonomous gatherings. For instance, you could utilize a Kruskal-Wallis H test to comprehend whether test execution, estimated on a nonstop scale from 0-100, contrasted in view of test nervousness levels (i.e., your reliant variable would be "test execution" and your free factor would be "test uneasiness level", which has three free gatherings: understudies with "low", "medium" and "high" test tension levels).

- **Assumptions of Kruskal- Wallis Test**

1. Your reliant variable ought to be estimated at the ordinal or constant level (i.e., stretch or proportion). Instances of ordinal factors incorporate Likert scales (e.g., a 7-point scale from "firmly concur" through to "emphatically dissent"), among alternate approaches to positioning classifications (e.g., a 3-point scale making sense of how much a client enjoyed an item, going from "Not without a doubt", to "It is alright", to "Indeed, a great deal"). Instances of consistent factors incorporate correction time (estimated in hours), knowledge (estimated utilising intelligence level score), test execution (estimated from 0 to 100), weight (estimated in kg, etc.
2. Your free factor ought to comprise at least two unmitigated, autonomous gatherings. Normally, a Kruskal-Wallis H test is utilised when you have at least three downright

free gatherings. Yet, it tends to be used for only two gatherings (i.e., a Mann-Whitney U test is all more generally used for two gatherings).

3. You ought to have freedom of perception, and that means that there is no connection between the perceptions in each gathering or between the actual gatherings. For instance, there should be various members in each meeting, and no member should be in more than one gathering.
4. To know how to decipher the outcomes from a Kruskal-Wallis H test, you need to decide if the disseminations in each gathering (i.e., the dispersion of scores for each gathering of the free factor) have a similar shape (which likewise implies a similar fluctuation). To comprehend what this means, investigate the chart below:



In the graph on the left, the circulation of scores for the "Caucasian," "African American," and "Hispanic" bunches has a similar shape. Then again, in the chart on the right, the appropriation of scores for each gathering is not indistinguishable (i.e., they have various shapes and change abilities).

13.5 Wilcoxon Signed Ranked Test

- **Meaning of Wilcoxon signed rank test**

The Wilcoxon signed-rank test is a nonparametric test identical to the reliant t-test. As the Wilcoxon marked rank test doesn't expect ordinariness in the information, it very well may be utilised when this suspicion has been disregarded, and the utilisation of the reliant t-test is improper. Contrasting two arrangements of scores that come from similar participants is used. This can happen when we wish to examine any adjustment of scores starting with one-time points and then moving on to the next or when people are exposed to more than one condition.

- **Steps to perform Wilcoxon signed rank test**

1. State Your Hypotheses:

- Invalid Speculation (H0): This speculation recommends that there's no distinction between the matched perceptions in the populace. In easier terms, the change between the two estimations (frequently meant "d") for each set of information focuses is zero across the populace.
- Elective Speculation (H1): This theory expresses something contrary to the invalid theory. It proposes there's a genuinely huge distinction between the matched perceptions. This distinction could be one-followed (coordinated, anticipating a shift in a particular course) or two-followed (undirected, expecting a shift in one or the other course).

2. Calculate the Differences: For each set of data of interest in your dataset, register the contrast between the two qualities. This distinction addresses the change you're keen on dissecting (e.g., the contrast in test scores when a mediation).

3. Rank the Distinctions (Overlooking Signs): Center exclusively around the outright upsides of the distinctions you determined in sync 2. Dismiss the positive or negative finishes paperwork for the time being. Rank these outright qualities from the littlest (rank 1) to the biggest (most elevated rank). In the event that you experience ties (various contrasts with a similar outright worth), relegate them to a typical position (e.g., two contrasts tied for the third spot would both be doled out a position of 3.5).

4. Reintroduce Signs and Dole out Marked Ranks: Presently, return to your unique distinctions (with signs) and coordinate them with their comparing positions from stage 3. Assuming the first contrast was positive, the marked position is basically the doled-out rank from stage 3. In any case, considering the first distinction was negative, duplicate the relegated rank by - 1 to get the marked position.

5. Calculate the W Statistic: There are really two forms of the W measurement you can utilise, contingent upon the indications of your disparities:

- W+: Aggregate the marked positions for every one of the positive distinctions.
- W-: Total the marked positions for every one of the negative distinctions.

6. Determine the Basic Value: Counsel a Wilcoxon Marked Rank basic worth table. This table will give basic qualities to various example sizes (n) and importance levels (as a rule indicated by alpha, frequently set at 0.05).

7. Interpret the Results: Think about the outright worth of your W measurement (either W+ or W-) to the basic worth obtained from the table. Assuming the outright

cost of your W measurement is not exactly or equivalent to the basic worth, you can dismiss the invalid speculation. This proposes a genuinely tremendous contrast between the matched perceptions.

- **Knowledge Check 2**

  State True or False

  1. Normally, a Kruskal-Wallis H test is utilised when you have at least three downright free gatherings, yet it tends to be used for only two gatherings. (True)
  2. It is vital to understand that the Kruskal-Wallis H test is an omnibus test measurement and can't let you know which explicit gatherings of your dependent factor are genuinely essentially not quite the same as one another. (False)
  3. You want to do this since it is possibly suitable to utilise a Wilcoxon marked rank test if your information "passes" three suppositions that are expected for a Wilcoxon marked rank test to give you a legitimate outcome. (True)
  4. Assuming the outright worth of your W measurement is exactly or equivalent to the basic cost, you can dismiss the invalid speculation. (False)

- **Outcome-Based Activity 2**

  A scientist concentrates on the viability of another rest-preparing program. They observed the rest length of the 10 newborn children when partaking in the program. Could you at any point utilise a Wilcoxon Signed Rank test to break down this information and decide whether the rest preparing program fundamentally affects newborn child rest span? Make sense of your reasoning for picking this test and what the signs in the ranked differences address.

13.6 Summary

- Non-parametric analysis is utilised when there is slanted information, and it includes procedures that don't rely upon information relating to a specific conveyance.
- Whenever a couple of presumptions in the given populace are unsure, we utilise non-parametric tests, which are likewise viewed as parametric partners.

- The objective of this test is to recognise whether a difference between genuine and anticipated information is because of possibility or a connection between the viable factors.
- The levels of opportunity in a measurable estimation address the number of factors that can shift in a computation.

13.7 Keywords

- **Non-parametric tests—Non-parametric tests are the numerical techniques utilised in factual speculation testing, which don't make presumptions about the recurrence conveyance of the factors** to be assessed.
- **Chi-square test—The Chi-square test is a measurable system for determining the contrast between** noticed and anticipated information.
- **Mann Whitney U Test**- The Mann-Whitney U test is a non-parametric test, meaning it doesn't depend on suspicions about the fundamental information conveyance (like ordinariness).
- **Kruskal Wallis Test**- The Kruskal-Wallis H test (at times likewise called the "one-way ANOVA on positions") is a position-based nonparametric test that can be utilised to decide whether there are genuinely massive contrasts between at least two gatherings of a free factor on a consistent or ordinal ward variable.
- **Wilcoxon signed Rank Test**- As the Wilcoxon marked rank test doesn't expect ordinariness in the information, it very well may be utilised when this suspicion has been disregarded. The utilisation of the reliant t-test is improper.

13.8 Self-Assessment Questions

1. What are the major differences between parametric tests and non-parametric tests?
2. What are the steps involved in calculation by the Chi-Square Test?
3. What is the formula used for Mann Whitney U Test?
4. What are the major assumptions of the Kruskal Wallis Test?
5. What do you mean by the Wilcoxon signed rank test?

13.9 References/ Reference Reading

- Gibbons, Jean Dickinson, and Subhabrata Chakraborti. Nonparametric Statistical Inference. 6th ed., CRC Press, 2020.

- Das, Rituparna, and Subhasish Das. Applied Statistical Methods: Nonparametric Tests and Multivariate Analysis. Narosa Publishing House, 2018.
- Siegel, Sidney, and N. John Castellan. Nonparametric Statistics for the Behavioral Sciences. 2nd ed., McGraw-Hill, 1988.
- Puri, Madan L., and Pranab K. Sen. Nonparametric Methods in Multivariate Analysis. Wiley, 2021.
- Agarwal, B. L. Basic Statistics. 6th ed., New Age International Publishers, 2016.

Unit 14- Multivariate analysis

Learning Outcomes

- Students will be able to define multivariate analysis
- Students will be able to evaluate the steps to perform factor analysis
- Students will be able to analyses the methods of clustering
- Students will be able to understand the merits and demerits of principal component analysis
- Students will be able to remember the types of discriminant analysis

Structure

14.1 Introduction to Multivariate Analysis

- **Meaning of Multivariate analysis**

Customarily, the measurable investigation is frequently centred around the connection between two factors (bivariate examination) or the normal of a solitary variable across various gatherings. Multivariate analysis breaks liberated from this constraint. It permits us to dig into the complicated existence where various factors impact a peculiarity all the while. The multivariate analysis goes past the constraints of concentrating on single factors, offering a diverse focal point to look at how numerous elements collaborate and impact certifiable situations.

- **Importance of multivariate analysis**

Multivariate analysis is a unique advantage in the realm of information investigation. Dissimilar to customary strategies that pay attention to single factors, it dives into the mind-boggling reality where numerous elements entwine. The following are three key justifications for why the multivariate examination is so significant:

1. **Unveiling Stowed away Examples and Relationships:**
   Envision concentrating on the variables affecting lodging costs. While area could appear to be a significant determinant, different factors like neighbourhood well-being, school locale quality, and nearness to conveniences additionally assume a part.
2. **Improved Independent direction and Asset Allocation:**
   In the business world, understanding client conduct is vital to progress. Multivariate examination assists organisations with dissecting various elements like socioeconomics, buy history, and online ways of behaving to recognise particular client portions.
3. **Enhanced Comprehension of Circumstances and logical results Relationships:**
   While the connection is not guaranteed to approach causation, multivariate investigation permits us to dig further into the "why" behind noticed connections. By genuinely controlling for the impact of different factors, we can detach the genuine effect of a particular element on the result of interest.

14.2 Factor analysis

- **Meaning of factor analysis**

Factor analysis, a technique inside the domain of insights and part of the general linear model (GLM), effectively gathers various factors into a more modest arrangement of variables. It catches the greatest common difference among the factors and gathers them into a brought-together score, which can consequently be used for additional analysis. Factor examination works under a few presumptions: linearity in connections, nonappearance of multicollinearity among factors, consideration of significant factors in the examination, and veritable relationships among' s factors and factors.

- **Steps to perform factor analysis**

Factor analysis is a measurable strategy used to portray fluctuation among noticed, corresponding factors as far as a possibly lower number of unnoticed factors called factors. Here are the general advances engaged with leading a component examination:

1. **Decide the Reasonableness of Information for Component Examination**

- Bartlett's Test: Check the importance level to decide whether the connection lattice is reasonable for factor investigation.
- Kaiser-Meyer-Olkin (KMO) Measure: Confirm the examining sufficiency. A worth more noteworthy than 0.6 is, for the most part, viewed as OK.

2. **Pick the Extraction Technique**

- Principal Component Analysis (PCA): Utilised when the fundamental objective is information decrease.
- Principal analysis of frequencies (PAF): Utilised when the primary objective is to distinguish hidden factors.

3. **Factor Extraction**

- Utilize the picked extraction strategy to recognise the underlying elements.
- Extricate eigenvalues to decide the quantity of variables to hold. Factors with eigenvalues more noteworthy than 1 are ordinarily held in the examination.
- Register the underlying component loadings.

4. **Decide the Quantity of Elements to Hold**

- Scree Plot: Plot the eigenvalues in plummeting request to picture where the plot levels off (the "elbow") to decide the number of variables to hold.
- Eigenvalues: Hold factors with eigenvalues more prominent than 1.

5. **Factor Revolution**

- Symmetrical Revolution (Varimax, Quartimax): Expects that the variables are uncorrelated.

- Angled Turn (Promax, Oblimin): Permits the elements to be corresponded.
- Pivot the variables to accomplish a less complex and more interpretable component structure.
- Analyze the turned element loadings.

6. Decipher and Mark the Elements

- Investigate the turned variable loadings to decipher the basic importance of each element.
- Allocate significant marks to each factor in view of the factors with high loadings on that variable.

7. Figure Component Scores (if necessary)

- Ascertain the element scores for every person to address their worth on each variable.

8. Report and Approve the Outcomes

- Report the last component structure, including factor loadings and communalities.
- Approve the outcomes utilising extra information or by directing a corroborative component examination if important.

14.3 Cluster analysis

- **Meaning of cluster analysis**

Cluster analysis, otherwise called clustering, is a technique for information mining that bunches comparable information focuses together. The objective of bunch examination is to partition a dataset into gatherings (or bunches) to such an extent that the data of interest inside each gathering are more like each other than the data of interest in different gatherings. This cycle is frequently utilised for exploratory information examination and can assist with recognising examples or connections inside the information that may not be quickly self-evident. There are various calculations utilised for bunch examination, like k-implies, progressive grouping, and thickness-based bunching. The decision to calculate will rely upon the particular necessities of the examination and the idea of the information being dissected.

- **Methods of Clustering**

1. **Partitioning Method:** It is utilised to make segments of the information to frame bunches. On the off chance that "n" segments are finished on "p" objects

of the data set, each segment is addressed by a bunch and n < p. The two circumstances that should be considered with this Dividing Grouping Technique are as follows: One goal ought to have a place with just a single gathering. There ought to be no gathering without even a solitary reason.

2. **Hierarchical Method:** In this technique, a progressive disintegration of the given arrangement of information objects is made. We can group progressive techniques and will actually want to know the motivation behind characterisation based on how the various levelled decay is shaped. There are two sorts of approaches for the making of various levelled disintegration, they are:
3. **Agglomerative Methodology:** The agglomerative methodology is otherwise called the granular perspective. At first, the given information is partitioned into which items structure separate gatherings. From there on, it continues to blend the articles or the gatherings that are near each other, which implies that they show comparable properties. This combining system goes on until the end condition holds.
4. **Divisive Methodology:** The disruptive methodology, otherwise called the hierarchical methodology, begins with the information protests of a similar group. Then, individual groups are partitioned into little bunches by a persistent cycle. The cycle goes on until the state of end is met or until each bunch contains one item.

The two methodologies which can be utilised further to develop the Various levelled Grouping Quality in Information Mining are: -

- One ought to painstakingly examine the linkages of the article at each parcelling of progressive grouping.
- One can involve various levels of agglomerative calculation for the combination of progressive accumulation. In this methodology, first, the articles are gathered into miniature bunches.

5. **Density-Based method: The thickness-put-together strategy chiefly centres on** respect to thickness. In this strategy, the given bunch will continue to develop ceaselessly as long as the thickness in the area surpasses some edge, i.e., for every data of interest inside a given group.

6. **Grid-Based method: In the Matrix-Based strategy, a framework is shaped by utilising the items together. The article space is quantised into a limited number of cells that form a lattice structure. One of the significant benefits of the matrix-based technique is its quick handling time, and it is reliant just on the number of cells in each aspect of** the quantised space.
7. **Model-Based method:** In the model-based technique, every one of the groups is speculated to find the information which is the most ideal for the model. The grouping of the thickness capability is utilised to find the bunches for a given model
8. **Constraint-based method: The limitation-based grouping technique is performed by the fuse of use or client-situated** requirements. A requirement alludes to the client's assumption or the properties of the ideal bunching results.

- **Knowledge check-1**

  Fill in the blanks

  1. By breaking down numerous factors together, _____methods can uncover stowed-away examples and connections that may be missed by checking every variable in disengagement out. (**Multivariate**/ Variance)
  2. _____analysis is a measurable strategy used to portray fluctuation among noticed, corresponded factors as far as a possibly lower number of unnoticed factors called factors. (Multivariate/ **Factor**)
  3. There are various calculations utilised for _____, like k-implies, progressive grouping, and thickness-based bunching. (**Clustering**/ discriminant)
  4. _____examiners utilize multivariate examination to evaluate venture chance and assemble ideal portfolios. (Frequency/ **Monetary**)

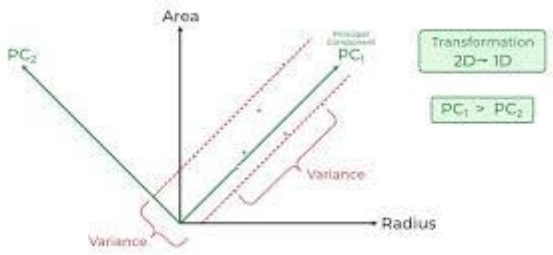
- **Outcome-based activity-1**

  Imagine you as a showcasing supervisor for a dressing store with a huge client dataset. Every information point incorporates data like age, buy history, and favoured styles. How might you utilize cluster analysis to distinguish particular customer groups with comparative purchasing propensities? Make sense of the reasoning you would take and the insights you hope to acquire.

14.4 Principal Component Analysis

- **Meaning of Principal Component Analysis**

The Principal Component Analysis(PCA) strategy was presented by the mathematician Karl Pearson in 1901. It deals with the condition that while the information in a higher layered space is planned to information in a lower aspect space, the fluctuation of the information in the lower layered space ought to be greatest. Principal Component (PCA) is a factual technique that utilises a symmetrical change that switches a bunch of connected factors over completely to a bunch of uncorrelated variables. CA is the most generally involved device in exploratory information examination and in AI for prescient models. Additionally, Principal Component Analysis (PCA) is a solo learning calculation strategy used to inspect the interrelations among a bunch of factors. It is otherwise called a general component examination where relapse decides a line of best fit.



1. Principal Component Analysis (PCA) is a method for decreasing dimensionality that distinguishes a bunch of symmetrical tomahawks, called head parts, that catch the most extreme change in the information.
2. The first principal part captures the most variety in the information; however, the second head part captures the greatest fluctuation that is symmetrical to the main head part, etc.
3. Principal Component Analysis can be utilised for different purposes, including information representation, highlight determination, and information pressure. In information perception, PCA can be used to plot high-layered information in a few aspects, making it more straightforward to decipher.

4. In Principal Component Analysis, the data is expected to be conveyed in the difference of the highlights; that is, the higher the variety in a component, the more data that elements convey.

- **Advantages and disadvantages of principal component analysis**

Advantages

1. Dimensionality Decrease: Principal component analysis is a famous method utilised for dimensionality decrease, which is the most common way of lessening the number of factors in a dataset.
2. Highlight Determination: Principal Component Analysis can be utilised to include choice, which is the most common way of choosing the main factors in a dataset.
3. Information Representation: Principal Component Analysis can be utilised for information perception. By lessening the number of factors, PCA can plot high-layered information in a few aspects, making it simpler to decipher.
4. Multicollinearity: Principal Component analysis can be utilised to manage multicollinearity, which is a typical issue in a relapse investigation where at least two free factors are profoundly corresponded.

Demerits

1. Understanding of Principal Components: The key parts of Principal Component Analysis are direct blends of the first factors, and interpreting them as far as the first variables are frequently troublesome.
2. Information Scaling: Principal component analysis is delicate to the size of the information. If the information isn't scaled as expected, then PCA may not function admirably.
3. Data Misfortune: Principal component analysis can cause data misfortune. While it diminishes the number of factors, it can likewise prompt data loss. The level of data misfortune depends upon the number of head parts chosen.
4. Non-straight Connections: Principal component analysis expects that the connections between factors are direct. Be that as it may, in the event that there are non-straight connections between factors, Principal component analysis may not function admirably.

14.5 Discriminant Analysis

- **Meaning of discriminant analysis**

Discriminant analysis (DA) is a multivariate method which is used to partition at least two gatherings of perceptions (people) based on factors estimated on each trial unit (test) and to find the effect of every boundary in separating the gatherings. What's more, the expectation or distribution of recently characterised perceptions to recently indicated gatherings might be inspected involving a direct or quadratic capability for relegating every person to existing gatherings. This should be possible by figuring out which bunch every individual has a place with.

- **Types of Discriminant Analysis**

1. Linear Discriminant analysis

Frequently known as LDA, it is a regulated methodology that endeavours to foresee the class of the Reliant Variable by using the direct mix of the Free Factors.

2. Quadratic Discriminant analysis

It is a subtype of linear Discriminant analysis (LDA) that utilises quadratic mixes of free factors to foresee the class of the reliant variable.

- **Knowledge Check-2**

State True or False

1. Principal Component Analysis (PCA) is a method for decreasing dimensionality that distinguishes a bunch of symmetrical tomahawks, called head parts, that catch the most extreme change in the information. (**True**/False)
2. Frequently known as QDA, it is a regulated methodology that endeavours to foresee the class of the Reliant Variable by using the direct mix of the Free Factors. (True/ **False**)
3. Tests should be liberated from each other and autonomous. (**True**/ False)
4. Principal Component analysis cannot be utilised to manage multicollinearity, which is a typical issue in a relapse investigation where at least two free factors are profoundly corresponded. (True/ **False**)

- **Outcome-based activity-2**

You are filling in as an information expert for a bank. The bank needs to characterise advanced candidates into two groups: high credit hazard and low credit risk. You have

information on different variables like pay, financial assessment, and credit sum. How might you utilize discriminant analysis to accomplish this objective? Make sense of the steps engaged in building the model and deciphering the outcomes.

14.6 Summary

- It permits us to dig into the complicated existence where various factors impact a peculiarity all the while.
- Multivariate examination assists us with grasping this mind-boggling exchange and recognize the main supporters of understudy achievement.
- Customarily, the hazard may be passed judgment on in view of a solitary element like verifiable stock cost execution.
- Factor examination works under a few presumptions: linearity in connections, nonappearance of multicollinearity among factors, consideration of significant factors in the examination, and veritable relationships among' s factors and factors.
- The grouping results ought to be interpretable, understandable, and usable.
- One ought to painstakingly examine the linkages of the article at each parcelling of progressive grouping.
- In the Matrix-Based strategy, a framework is shaped by utilising the items together. The article space is quantised into a limited number of cells that form a lattice structure.

14.7 Keywords

- **Multivariate analysis**- Multivariate examination assists us with grasping this mind-boggling exchange and recognise the main supporters of understudy achievement.
- **Factor analysis**—Factor analysis, a technique within the domain of insights and part of the general linear model (GLM), effectively gathers various factors into a more modest arrangement of variables.
- **Homoscedasticity**- The difference of the factors ought to be generally equivalent across various levels of the variables.
- **Partitioning method**- On the off chance that "n" segments are finished on "p" objects of the data set, each segment is addressed by a bunch and n < p.

- **Density-based method**- In this strategy, the given bunch will continue to develop ceaselessly as long as the thickness in the area surpasses some edge, i.e., for every data of interest inside a given group.
- **Grid-based method—In the Matrix-based strategy, a framework is shaped by utilising the items together; the article space is quantised into a limited number of cells that form** a lattice structure.
- **Constraint-based method**—The limitation-based grouping technique is performed by combining use or client-situated requirements.
- **Principal Component Analysis**—Principal Component (PCA) is a factual technique that utilizes a symmetrical change to switch a set of connected factors completely to a set of uncorrelated variables.
- **Discriminant Analysis**—Discriminant analysis (DA) is a multivariate method used to partition at least two groups of perceptions (people) based on factors estimated on each trial unit (test) and to find the effect of every boundary in separating the groups.

14.8 Self-Assessment questions

1. What do you mean by multivariate analysis?
2. What are the steps to perform factor analysis?
3. What are the different methods of clustering?
4. What are the merits and demerits of principal component analysis?
5. What are the 2 main types of discriminant analysis?

14.9 References/ Reference Reading

- Johnson, Richard A., and Dean W. Wichern. *Applied Multivariate Statistical Analysis*. 6th ed., Pearson, 2018.
- Hair, Joseph F., et al. *Multivariate Data Analysis*. 8th ed., Cengage, 2019.
- Tabachnick, Barbara G., and Linda S. Fidell. *Using Multivariate Statistics*. 7th ed., Pearson, 2019.
- Sharma, Subhash. *Applied Multivariate Techniques*. Wiley India, 2020.
- Ramachandran, R., and R. Srinivasan. *Multivariate Statistical Analysis*. Prentice Hall India, 2021.

Unit 15- Statistical Quality Control

Learning Outcomes

- Students will be able to define the meaning of statistical quality control
- Students will be able to evaluate the types of control charts for variables
- Students will be able to analyses the types of control charts for attributes
- Students will be able to understand the importance of process capability analysis
- Students will be able to remember the principles of the six sigma methodology

Structure

15.1 Concept of Statistical Quality Control

- **Meaning of statistical quality control**

Statistical quality control (SQC) is a foundation of present-day quality administration, using the force of insights to screen, investigate, and further develop creation processes. A proactive methodology goes past just reviewing completed items; all things being equal, SQC centres around ceaselessly checking the actual cycles to guarantee they reliably create top-notch yields. SQC use a set-up of measurable devices and strategies to accomplish this goal. A key component is statistical process control (SPC), which uses control graphs to portray process conduct after some time outwardly. These outlines lay out upper and lower control limits in light of authentic information, featuring any critical deviations that could show hidden issues.

- **Importance of Statistical Quality Control**

Statistical quality control (SQC) is a foundation of modern production, employing the force of measurements to accomplish three basic targets: indicators of process solidity, practices related to early identification of imperfections, and steady progress. All of these features perform in unison to enhance item quality, lower costs, and, last but not least, the consumer retention rate.

1. **Assuring Cycle Stability**—At the centre of the table, the Supporting Quality Cycle (SQC) is the gauge of interaction extent. Using a control outline, for instance, SQC continually monitors the development process, observing oddities and addressing them where necessary.

2. **Differential and Correction at the Initial Stage**- This is another significant element that marks the importance of SQC as it can identify and correct quality problems right at the initial stages of the creation cycle. With such devices as acknowledgement inspecting, the makers are in a position to detect the non-adjusting gadgets a long time before the actual end of production. Such restrictions demoralize and retain sub-par products from reaching customers.

3. **Continual Change**—SQC is not just a dormant supervising tool; it is a catalyst for continuing improvement. From the rehearsals, information accumulates concerning valuable experiences on how the processes are executed under SQC. Decision makers

are able to identify disruptions, if not blockages or failures, and possibly the origin of variability that makes up the process.

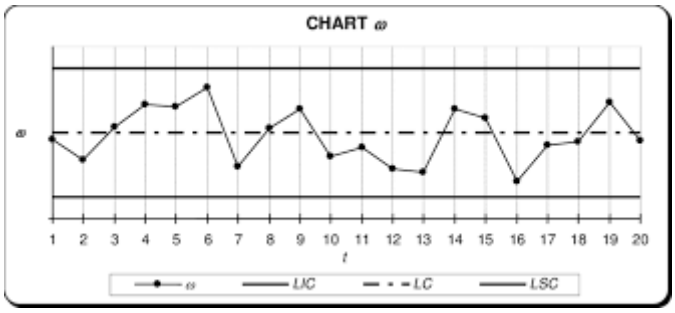
15.2 Control Charts for Variables

- **Meaning of control chart for variables**

Inside the domain of statistical quality control (SQC), control charts for variables play an essential role in defending cycle security. These diagrams are graphical devices explicitly intended to screen ceaseless information, giving continuous insights into an interaction's focal propensity and changeability.

The centre construction of a control chart for variable comprises three key components:

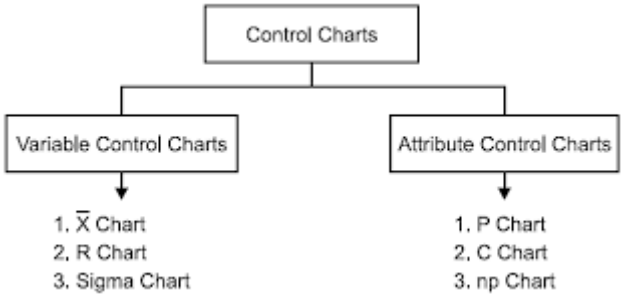
1. Centre Line: This line addresses the cycle's typical presentation, commonly determined as the mean of authentic information. It serves as a benchmark against which ensuing estimations are evaluated.
2. Control Limits: These are measurably inferred upper and lower limits, put a particular distance (frequently a few standard deviations) from the middle line. These cutoff points characterise the normal scope of variety intrinsic in a steady cycle.
3. Sample Statistics: The real estimations, frequently gathered into subgroups for improved examination, are plotted on the graph. These focuses uncover the cycle's conduct over the long run and feature any possible deviations from the normal example.



Types of control charts for variables

Control charts are the workhorses of statistical quality control (SQC) for observing cycles that create ceaseless, quantifiable information. These diagrams portray a cycle's focal inclination (normal) and fluctuation over the long haul, permitting us to recognise

any deviations from the wanted execution. How about we dig into the most widely recognised control diagrams for factors:



1. X̿-R Chart: This powerful couple consolidates two diagrams: the X̿ chart which tracks the mean (average) of subgroups (tests) taken at standard stretches, while the R chart which screens the range (distinction between the most noteworthy and least value) inside every subset. This mix gives an all-encompassing perspective on the cycle's focal inclination and inconstancy. The X̿ graph identifies shifts in the normal, while the R diagram uncovers changes in the spread of the information.
2. X̿-S Chart: Like the X̿-R outline, the X̿-S graph tracks the mean yet uses the standard deviation (S) as a proportion of inconstancy. The standard deviation offers an all the more measurably strong proportion of spread contrasted with the reach, particularly for bigger subgroup sizes. This outline is especially valuable when the subgroup size is steady.
3. Individual (X) and Moving Reach (MR) Chart: This mix centres around individual pieces of information close by their nearby fluctuation. The X chart plots every individual estimation, taking into consideration the representation of patterns and exceptions. The MR chart computes the reach between back-to-back pieces of information inside subgroups, giving experiences into momentary cycle inconstancy.
4. Moving Average (MA) and Moving Reach (MR) Chart: This approach utilizes moving midpoints to streamline the intrinsic arbitrariness in individual estimations. The MA chart plots the normal of a predefined number of late subgroups featuring longer-term patterns in the process focal propensity. The MR chart still screens the reach inside subgroups, giving data on transient fluctuation even with the streamlined normal.

15.3 Control Charts for Attributes

- **Meaning of control charts for attributes**

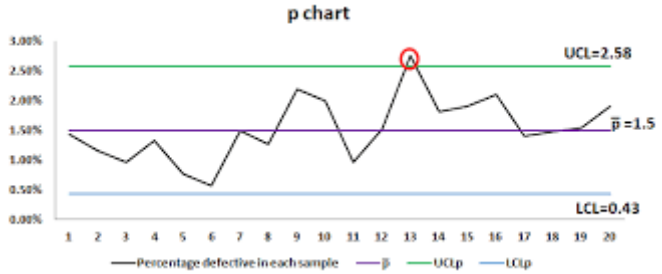
Within the focus of textual quality control, statistical quality control (SQC), control charts for attributes play an important role in monitoring cycles that generate nonnumerical data. Unlike control charts for variables that involve controlling characteristics such as amount, length, or weight, trait charts work based on qualities that can be either controlling or non-controlling, defective or non-defective.

The power of control charts for attributes lies in the ability of process conduct to be outward noticeable. While these devices superimpose what appears to have been genuinely calculated upper and lower control limits (UCL and LCL) on the graph, one gets a clear picture of what constitutes averaged cycle variation. Any amount of information focused externally on these cutoff points signify possible troubles that ought to be considered.

- **Types of control charts for attributes**

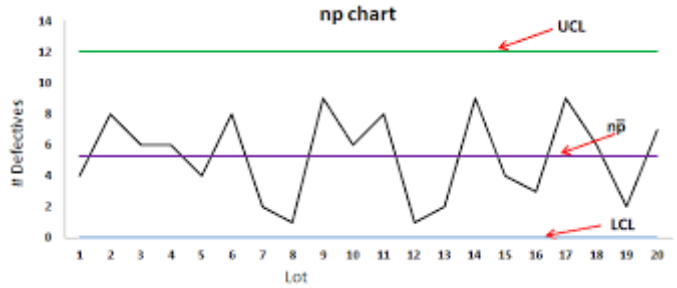
Through SPC, one is able to meet the set standardised quality thereby playing a significant role in maintaining product quality. While control outlines for factors define things like weight or aspect mathematically, processes can contain credits – attributes that are summarised as adjusting or non-adjusting. Here, we dive into the three essential control graphs utilised for trait information: Here, we dive into the three essential control graphs utilised for trait information:

1. **p-graph (extent nonconforming):** This outline represents the progress of an instance with the damaged units in fractions. That is good for situations where the sizes for the examples remain homogenised. The p-graph filters the interaction's ability to provide a certain level of reliability when modifying things through the interaction's range. It employs binomial appropriation as a means of estimating control limits a bit that contains any important deviations that can point to shifts at the same time.



1. **np-graph (number nonconforming):** When test size changes, the np-outline becomes a significant instrument. It centers around the actual number of

inadequate units inside each example instead of the extent. The np-diagram utilizes the Poisson circulation to lay out control limits, which change naturally based on the differing test sizes. This takes into account powerful cycle observing, paying little heed to test size varieties.



2. **c-chart (number of deformities per unit):** This graph takes special care of situations where various imperfections can happen on a solitary unit. It tracks the total number of imperfections saw inside each example unit. The c-outline uses the Poisson dissemination, like the np-graph, to ascertain control limits for the normal number of deformities per unit. This approach distinguishes circumstances where the general imperfection rate may be expanding, regardless of whether the extent of faulty units stays stable.

- **Knowledge Check-1**

  Fill in the blanks

  1. _____(SQC) is a foundation of present-day quality administration, using the force of insights to screen, investigate, and further develop creation processes. (**Statistical quality control**/ Statistical variable control)
  2. The force of _____lies in their capacity to separate between innate, irregular cycle variety and assignable reasons for variety.(Frequency tables/ **control charts**)
  3. Like the X̄-R outline, the_____ graph tracks the mean yet uses the standard deviation (S) as a proportion of inconstancy.(**X̄-S**/ XT)
  4. Inside the domain of statistical quality control (SQC), control charts for _____serve an imperative job in observing cycles that produce subjective information.(Variables/ **attributes**)

- **Outcome based activity -1**

An assembling organisation tracks the everyday pace of deficient lights created on a p-diagram. As far as possible haven't been penetrated in weeks. Out of nowhere, a point falls outside the upper control limit. Make sense of the potential purposes behind this point and the means the organisation ought to take accordingly.

15.4 Process Capability Analysis

- **Meaning of process capability analysis**

Process capability analysis (PCA) is a factual method utilised in assembling and other quality-driven ventures to evaluate a cycle's capacity to deliver yields that meet pre-characterised details reliably. It digs further than basically assessing individual items; PCA assesses the inborn changeability of the whole cycle itself. PCA use factual instruments to look at the regular spread of a cycle's results, measured by its standard deviation, to the laid out upper and lower determination limits (USL and LSL) for a specific trademark. This correlation is regularly communicated through capability indices, like Cp and Cap. These lists give a mathematical portrayal of how very much focused and how fit an interaction is of meeting particulars.

- **Importance of process capability analysis**

Process capability analysis (PCA) remains as a fundamental instrument inside the munitions stockpile of factual quality control (SQC). It dives further than just guaranteeing a cycle stick to details; PCA evaluates an interaction's innate capacity to create yields that meet those determinations reliably. This qualification is critical, as a cycle sticking to particulars every so often may not ensure steady quality over the long haul. PCA use measurable methods to evaluate the inborn changeability of an interaction. This inconstancy, frequently addressed by the standard deviation, mirrors the normal changes present inside the cycle. PCA then looks at this inconstancy to the laid out upper and lower particular cutoff points (USL and LSL) for the item or administration being created. By working out measurements like Cp and Cpk, PCA gives an unmistakable image of the cycle's capacity. A high Cp shows the cycle can possibly create well inside as far as possible, while a high Cpk implies the interaction is right now focused inside those cutoff points, limiting the probability of deformities. The significance of PCA lies in its capacity to:

1. Foresee future performance: By understanding the inborn changeability of a cycle, PCA permits makers to foresee with more noteworthy certainty the probability of delivering adjusting yields.
2. Recognize improvement opportunities: A low Cpk esteem proposes the cycle isn't in every case meeting determination. PCA assists pinpoint the root with causing of this fluctuation, taking into account designated intercessions to further develop process solidness and lessen scrap rates.
3. Decrease costs: By limiting the creation of flawed items, PCA prompts tremendous expense investment funds. This incorporates diminishing material waste, limiting revamp endeavors, and bringing down guarantee claims. Also, by streamlining process proficiency, PCA can add to expanded creation yield.

15.5 Six Sigma Methodology

- **Meaning of six sigma methodology**

Six sigma is an information driven system designed to upset business process improvement. It goes beyond the issues of quality assurance and gets to the point of perfecting deformities, let alone differentiating varieties. S Six Sigma was created at Motorola throughout the 1980s; it has advanced to become a universal standard of practical excellence in multiple organisations. The third guideline of Six Sigma involves the use of a factual approach. About it characterizes a “sigma” as numerical difference from the mean. For instance, a Six Sigma process boasts of a deformity rate a simple DPMO stands for defects per million opportunity and, in this particular case, it occurred at a DPMO, which also points to a remarkably high level of value. This means perfect ideal objects and services, which meet and even exceed the customer expectations, which in turn spur continuous enhancement.

Six Sigma accomplishes this exceptional accomplishment through an organised methodology known as DMAIC:Six Sigma accomplishes this exceptional accomplishment through an organised methodology known as DMAIC:

1. Define: It properly sets a perception of the issue or an opportunity for growth during this stage. It comprises of aspects such as defining the client requirements, constraints of the process, and desired outcomes.

2. Measure: It moves to assessing the current state of the interaction and this marks the center of the movements. Sorting and analysis turn into basic, initiating major run indicators (KPIs) to check distortions and variations.

3. Analyze: The preparation phase in this stage provides main driver investigation. Measures and approaches are used to convey which factors contribute to the defects and interaction discrepancies.

4. Improve: This stage continues the fragments of knowledge that emerged from the examination. It involves rinsing answers to extract the underlying causes. It might mean updating or changing processes, reparations of gears, or even worker training.

5. Control: The final stage ensures that the enhanced changes made in an organisation are safeguarded for future continuity—vigilance or monitoring and correcting parts to prevent backsliding and ensure that the relationship remains optimally healthy.

- **Principles of Six Sigma methodology**

Six Sigma is an information-based methodology that is appreciated for its profitability in enhancing cycle unswerving nature and usefulness. It settles upon five centre rules that guide its execution and fuel its prosperity: It settles upon five centre rules that guide its execution and fuel its prosperity:

1. Customer Focus: The cutting-edge Six Sigma frameworks concentrate on the client, as all enhancement attempts ensure optimal value delivery. This standard will underline understanding client needs, defining Kaizen or key-to-quality attributes (CTQ) that correspond to those problems, and using them as a measure to reach process improvement objectives.

2. Data-Driven Choice Making: This approach goes against the utilitarian approach, where instinct and mystery are considered waste in Six Sigma. All the stages, starting with the issue of distinguishable proof and ending with the execution of the arrangement, are totally based on information investigation.

3. DMAIC Cycle: Six Sigma operates in a systematic procedure that is called DMAIC, an acronym for Define, Measure, Analyze, Implement, and Control. This recurring pattern provides a guide to how a particular organisation should deal with progress projects. This system commences with characterising the issue and client necessities, which are Characterised, and then cautiously estimates the present system execution, which is measured. Dive down to the information of the breakdown stage to dissect the specifics of varieties of and deformities source. Based on such concerns, the

4. Continuous Improvement: It is very common to hear business professionals in organisations practicing Six Sigma talk about never-ending improvement. It is not merely a practice of rectifying an unfortunate event but rather a continuous process of changing one’s mindset.

5. Defect Prevention: Six Sigma wants to prevent absconds from even occurring and not merely recognise that they have happened down the line.

- **Knowledge Check-2**

  State True or False

  1. By understanding the inborn changeability of a cycle, PCA permits makers to foresee with more noteworthy certainty the probability of delivering adjusting yields. (**True**/ False)
  2. A high Cpk esteem proposes the cycle isn't in every case meeting determination. (True/ **False**)
  3. Six Sigma focuses on the client, guaranteeing all improvement endeavours line up with conveying the most extreme worth. (**True**/ False)
  4. Six Sigma doesn’t limit mistakes and imperfections, prompting cost decreases and further developing benefits. (True/**False**)

- **Outcome-based activity-2**

  You are the quality supervisor at an assembly office, encountering a high pace of imperfections in basic parts. Make sense of how you would carry out the Six Sigma DMAIC strategy to resolve this issue. Please make certain to frame the particular advances you would take in each stage (Characterize, Measure, Examine, Improve, Control) to distinguish the underlying driver of the imperfections and carry out an answer to guarantee their disposal.

15.6 Summary

- SQC use a set-up of measurable devices and strategies to accomplish this goal. A key component is statistical process control (SPC), which uses control graphs to portray process conduct after some time outwardly.
- It empowers early discovery and correction of value issues, forestalling the creation of non-adjusting items.
- Methods like gauge repeatability and Reproducibility (GR&R) studies evaluate this variety and survey, assuming that it fundamentally influences the capacity to recognize genuine interaction varieties.

- This powerful couple consolidates two diagrams: the $\bar{X}$ chart, which tracks the mean (average) of subgroups (tests) taken at standard stretches, and the R chart, which screens the range (the distinction between the most noteworthy and least value) inside every subgroup.
- The standard deviation offers an all the more measurably strong proportion of spread contrasted with the reach, particularly for bigger subgroup sizes.

15.7 Keywords

- **Statistical Quality Control**- Statistical quality control (SQC) is a foundation of present-day quality administration, using the force of insights to screen, investigate, and further develop creation processes.
- **Process Capacity Analysis**- This measurable method surveys how well a creation cycle can reliably meet predefined quality particulars.
- **Acceptance Sampling**—This strategy upgrades review endeavors by genuinely deciding the example size needed to make dependable assumptions about the nature of a larger group of things.
- **Control Charts for variables**—These diagrams are graphical devices explicitly intended to screen ceaseless information, giving continuous insight into the focal propensity and changeability of an interaction.
- **Sample statistics**- The real estimations, frequently gathered into subgroups for improved examination, are plotted on the graph.
- **$\bar{X}$-R Chart**- This powerful couple consolidates two diagrams: the $\bar{X}$ chart, which tracks the mean (average) of subgroups (tests) taken at standard stretches. In contrast, the R chart screens the range (the distinction between the most noteworthy and least value) inside every subset.
- **$\bar{X}$-S Chart**- Like the $\bar{X}$-R outline, the $\bar{X}$-S graph tracks the mean yet uses the standard deviation (S) as a proportion of inconstancy.
- **Individual (X) and Moving Reach (MR) Chart**—This mix centers around individual pieces of information grouped by their nearby fluctuation.
- **Control Charts for attributes**—In the domain of statistical quality control (SQC), control charts for attributes are imperative in observing cycles that produce subjective information.
- **p-graph**- This outline tracks the fraction of damaged units inside an example.

- **Np graph**- The np-diagram utilises the Poisson circulation to lay out control limits, which change naturally founded on the differing test sizes.
- **C- chart**- This graph takes special care of situations where various imperfections can happen on a solitary unit.
- **Process capability analysis (PCA)**- It is a factual method utilised in assembling and other quality-driven ventures to evaluate a cycle's capacity to deliver yields that meet pre-characterised details reliably.
- **Six Sigma Methodology-** Six Sigma is an information-driven system intended to improve business process improvement.

15.8 Self-Assessment questions

1. What do you mean by statistical quality control?
2. What are the different types of control charts for variables?
3. What are the different types of control charts for attributes?
4. What do you mean by process capability analysis?
5. What are the key principles of Six Sigma methodology?

15.9 References/ Reference Reading

- Montgomery, Douglas C. *Introduction to Statistical Quality Control*. 8th ed., Wiley, 2019.
- Gupta, J. P. *Statistical Quality Control: Using MINITAB, R, JMP and Python*. Wiley, 2023.
- Bedi, Suresh Kumar. *Quality Management*. Oxford University Press, 2021.
- Mahajan, M. *Statistical Quality Control*. Dhanpat Rai & Co., 2022.
- Mitra, Amitava. *Fundamentals of Quality Control and Improvement*. 4th ed., Wiley, 2021

Chapter 16- Decision Theory and Applications

Learning Outcomes

- Students will be able to define decision theory.
- Students will be able to evaluate the steps taken to reduce errors by making decisions under uncertainty.
- Students will be able to analyze the challenges posed by decision-making under risk.
- Students will be able to understand the significance of decision trees and payoff tables.
- Students will be able to remember applications of decision theory in business strategy.

Structure

16.1 Introduction to Decision Theory

- Meaning of decision theory
- Principles of decision theory
- Importance of decision theory

16.2 Decision-making under uncertainty

- Meaning of decision-making under uncertainty
- Challenges posed by decision-making under uncertainty
- Steps to reduce errors made by decisions under uncertainty

16.3 Decision-making under risk

- Meaning of decision-making under risk
- Challenges posed by decision-making under risk
- Knowledge check-1
- Outcome-based activity

16.4 Decision trees and Payoff tables

- Meaning of decision trees and payoff tables
- Elements of decision trees
- Significance of decision trees and payoff tables

16.5 Applications of decision theory in business strategy

- Knowledge check-2
- Outcome-based activity

16.6 Summary

16.7 Keywords

16.8 Self-assessment questions

16.9 References/ reference reading

16.1 Introduction to Decision Theory

- **Meaning of decision theory**

Decision theory, a connection of rationale, likelihood, and human way of behaving, digs into the fascinating universe of how decisions are made. It envelops two primary points of view: regularising and expressive. Regulating hypothesis recommends an ideal leader, using total data and impeccable estimations to go with ideal decisions. This glorified model fills in as a benchmark for assessing true direction. Distinct hypothesis recognises the messy truth of human decision. It investigates the inclinations, mental alternate ways (heuristics), and profound impacts that shape our choices, frequently driving us to go amiss from amazing sanity. Understanding these variables is urgent for determining how individuals really decide.

Decision theory outfits us with devices to examine these cycles. The utility hypothesis allows values to result, permitting the correlation of choices with various outcomes. Anticipated that worth considers probabilities and utilities should compute the typical result of a choice. Choice trees outwardly map out expected ways and their results, helping with clear correlation. At last, hazard avoidance recognizes our propensity to stay away from vulnerability, even at the expense of likely gains. By understanding these standards, people and associations can settle on additional educated and objective decisions, adjusting activities to objectives and hazard resistance.

- **Principles of decision theory**

Decision theory establishes a groundwork for settling on very much educated decisions, yet it perceives the intricacies of human behaviour. The following are five key rules that support this hypothesis:

1. Rationality: These standard structures are the centre of regulating choice hypothesis. It expects an entirely normal chief with complete information on every accessible choice, their related probabilities, and the likely results. This admired chief utilises faultless rationale and expands their normal utility by choosing the option with the best result.
2. Bounded Rationality: This illustrative hypothesis recognises the constraints of human cognisance. We don't always have complete data, and handling complex circumstances can be overwhelming. Limited soundness means that people frequently depend on alternate mental routes, known as heuristics, to work on direction. These heuristics can be proficient, but they may prompt inclinations that slant our decisions.
3. Expected Value: This standard assesses several potential outcomes of a decision. Instead, it involves the assumption-weighted mean of the qualities associated with every possible outcome, given the probability of each result. The regular worth of every option is the capability that chiefs can use to choose the strategy with the finest general ordinary worth.
4. Utility Theory: This structure acknowledges that people value different outcomes in a counterintuitive manner. The utility hypothesis allocates mathematical qualities (utilities) to distinct results based on the examination of choices which involve distinct results. Some might emphasise well-being rather than monetary thanks, which is seen in their downplayed utilities. Kahneman and Tversky's concept of expected utility is consequently taken in advance with the objective of expanding its pertinence to chiefs.
5. Risk Aversion: This standard acknowledges our inherent dislike of exposure based on the belief that nobody is comfortable being exposed. Even if a particular choice with a bigger potential chip also provides a bigger risk, many folks would rather go for a safer selection with fewer possible returns. Simply comprehending some of the aspects of hazard avoidance is necessary for predicting a dynamic way of behaving because sometimes exposure to the chosen course of action involves different rates of vulnerability.

6. These standards offer a compass for evaluating and advancing progressive current cycles. Although wonderful judiciousness could be an adored model, comprehending these central ideas assists individuals and associations in making fewer, possibly the most educated, and powerful decisions, even under intricate problems.

- **Importance of decision theory**

Decision theory is not an abstract area of academic interest only but has a number of well-developed concepts and tools to help look at the subtleties of decisions. Here's the reason why understanding the choice hypothesis is significant. Here's the reason why understanding the choice hypothesis is substantial:

1. **Improved Choice Making:** It provides people and associations with a systematic approach to appraise choices, which is known as the choice hypothesis. Chiefs can shift away from suspicion and instinct to different expectations, probability or likely outcomes, and other tendencies or inclinations (utilities). This systematic approach helps to manage the effects of tendencies or the use of heuristics, which leads to a better outcome.

2. **Risk Management:** Being defenceless is a fragment of being alive, and the choice hypothesis furnishes the procedures for managing risk so well. By converting probable outcomes into the likelihood of gains and losses and the value of their respective potential benefits, chiefs can gain insight into the expected negatives of every existing option and choose strategies that mitigate potential escalation of risk with risk aversion. This covers all well-planned actions and minimises the possible negative outcomes that might be a result of certain actions.

3**. Strategic Planning:** The choice hypothesis is a powerful tool when it comes to critical preparation. It permits associations to consider their key decisions crosswise over different time openings, including future inclinations, potential challenges and rival activities. By unveiling choice trees and finding out the probable values of such situations, the associations can make vital investments and decisions on the allocation of their resources better, increasing the likelihood of achieving the association's long-term goals and objectives.

4**. Negotiation and Struggle Resolution:** The Choice hypothesis participates in exchange and compromise. This way, it is possible to recognize score-related gamble propensities and requirements of various groups involved and develop cooperation

arrangements for mutual gain or enhance positions for bartering. It considers better exchanges and decreases the chance of consecutive fighting.

5. **Understanding Human Behavior:** Unlike other approaches, the choice hypothesis isn't only focused on predicting ideal decisions; it also provides a deep understanding of how people actually make decisions. Understanding mental urges and decision-making heuristics helps associations think about UIs, marketing approaches, and communications systems that guide people towards desired actions. This understanding of human behavioural patterns is critical in various disciplines encompassing the advertising industry and public arrangement.

In conclusion, the choice hypothesis asserts itself over speculative concepts. As a system for analysing choices, managing danger, and comprehending human conduct, it empowers individuals and associations to make progressively great decisions, ally with uncertainty, and achieve their goals.

16.2 Decision-making under uncertainty

- **Meaning of decision-making under uncertainty**

Decision-making under uncertainty is what is happening in our lives, happening at whatever point we should pick a game plan without complete information on the possible results. It is not at all like dynamic under conviction, where all prospects and their outcomes are known, where vulnerability tosses a cloak over what's in store. This absence of wonderful data can come from different variables, including:

- Restricted data: We may essentially not have sufficient data about the circumstance or the expected results of every decision.
- Inborn randomness: A few occasions are innately inconsistent, making it difficult to decide the specific outcomes of our activities.
- Complexity: Complex circumstances with various factors and associating elements can challenge exact forecasts.

In spite of these difficulties, the dynamic of vulnerability stays pivotal in different areas, from business speculations to individual life decisions. To explore this vagueness, people and associations utilise a scope of methodologies:

- Risk assessment: Via cautiously investigating the accessible data and recognising possible dangers, leaders can appraise the probability of different results and their related outcomes. This hazard appraisal illuminates decisions and possibly moderates expected adverse consequences.

- Situation planning: This approach includes considering various conceivable future situations, even those with low probabilities. By investigating a range of conceivable outcomes, chiefs can develop emergency plans and prepare for unforeseen events.
- Heuristics and biases: These enthusiastic other ways of allowing one's picked parts to drive poor decisions can, in the same manner, be gotten instruments when under threat. In this case, leaders can stay dry with a set of rules of thumb and prior experiences fast-forward and make decisions without feeling overwhelmed by the grand issues.

Management decision-making under conditions of risk involves balancing rationality in both the plans that will be implemented and the activities that will be carried out. Although achieving complete certainty may be challenging, a highly structured approach that considers readily available data, potential risks, and various other future scenarios encourages people and organisations to make rational decisions that minimise uncertainty and risk despite the fact that the future is always unpredictable.

- **Challenges posed by decision-making under uncertainty**

The concept of decision-making with great uncertainty, which is an ever-present experience in our lives, poses unique sets of challenges that are capable of changing the overall form of the decision-making process. Here is a more intensive glance at a portion of the key obstacles we face: Here is a more intensive glance at a portion of the key obstacles we face:

1. Restricted Information: Often, the information which contributes to the understanding of the forecasted outcome or the true nature of the situation stays veiled in the assessment of vulnerability. This absence of complete data makes it difficult to pinpoint the exact value and risk that goes with each option. There may be a situation where decision-makers rely on inadequate data, experience or probable assumptions, which leads to a decision of poor standards.

2. Mental Biases: It is since our minds are structured with other routes called heuristics, which help us analyse complex situations swiftly. However, these different ways can sometimes be misleading, at least on a number of occasions. Situational inclination, for example, mooring inclination, can cause us to overemphasise the primary point of information we receive, which may be misleading our analysis of several options. These impacts make it possible for us to look for apparent and good feedback that is

inconsistent evidence that we discard consistently. Both these propensities interfere with our reasoning ability and hinder our ability to seek rational action under threat.

3. Risk Aversion: Humans are, by default, risk-averse and will naturally gravitate towards choices with a higher degree of certainty no matter the risk/reward balance. This can, again, be quite a challenge, especially when dealing with vulnerability. It is perhaps reasonable to resist opting for choices that telegraph an exposure to unfortunate outcomes even though significant benefits could be reaped. This hazard avoidance can limit the ability to recognise and capitalise on those opportunities in order to advance the goals.

4. The Outlining Effect: We discuss how data can be introduced in ways that can either threaten or enhance our decision-making. Predisposition features, where the predisposition features construct an issue, may cause one to be inclined towards one decision rather than the other regardless of the basic information. This can make it difficult to follow objective decisions under vulnerability, as how a circumstance is presented can play a part in our decision-making processes.

5. The Scourge of Knowledge: If there is something that others require knowing, anticipating how they will feel or what they may see can be inconvenient. This "scourge of information" can arrest correspondence and alight on it, attempting to rationalize our choice-making cycle to other people, particularly when confronted by vulnerability. Its fluidity and sheer volume create such a scenario.

They point to difficulties that, if understood, can make us into better, careful leaders when vulnerable. Approaches such as social events, though as much information as could be conceived of, are prone to recognise potential biases and consider all the options, which can essentially help to address these challenges and proceed to more informed decisions.

- **Steps to reduce errors made by decisions under uncertainty**

Decision-making under uncertainty, where complete data is tricky, is a consistent human encounter. While amazing results are unthinkable, there are steps we can take to decrease mistakes in these circumstances essentially. Here is a breakdown of key methodologies:

1. **Gather as much data as possible:** The more information you have, the better you can assess likely results. Lead intensive exploration, talk with specialists, and consider assorted viewpoints. While complete sureness may be

unreachable, a powerful database fills in serious areas of strength for steady navigation.

2. **Identify and Relieve Biases:** We all harbour mental predispositions, mental easy routes that can prompt defective decisions. Monitoring these inclinations, for example, the tendency to look for predictable feedback (leaning toward data that affirms existing convictions) or mooring predisposition (overreliance on starting data), assists us with relieving their impact. Procedures like arguing for the sake of arguing or utilising post-mortem examination (envisioning the most pessimistic scenario situations) can uncover expected vulnerable sides.
3. **Embrace Situation Planning:** Don't restrict yourself to a solitary "best case" situation. Foster a scope of potential outcomes, taking into account both positive and adverse results. This permits you to expect possible barriers and foster alternate courses of action. Situation arranging cultivates a more reasonable comprehension of the choice scene and diminishes the gamble of being surprised by startling turns of events.
4. **Quantify Uncertainty:** Whenever the situation allows, use devices like likelihood hypothesis to allot probabilities to various results. This considers a more nuanced comprehension of the likely dangers and prizes related to every choice. Expected esteem computations, which think about the two probabilities and possible results, can assist and guide you towards decisions with the most noteworthy potential for progress.
5. **Seek Different Input:** Don't settle on choices in a vacuum. Talking with people from various foundations and with different aptitudes can open you to new viewpoints and possible vulnerable sides. Utilising the aggregate insight of a group can prompt more hearty and balanced choices, especially under states of vulnerability.
6. **Prepare to Adapt:** The truth of vulnerability is that things seldom unfold precisely as expected. Foster an adaptable methodology, considering course revision as new data arises. Consistently screen progress, be available to reconsider your system and embrace an eagerness to adjust as conditions direct.

16.3 Decision-making under risk

- **Meaning of decision-making under risk**

Dynamic risk consumes a particular space inside the more extensive domain of the choice hypothesis. It manages circumstances where we face different decisions, each with known expected results, yet the probability of every result is questionable. There's an unmistakable qualification among hazard and vulnerability: under risk, we can measure the probabilities of every result, regardless of whether they're not ensured. Imagine you're flipping a coin to choose which film to watch. You know there's a half opportunity of heads (one film) and a chance for half of tails (the other movie). The future result (heads or tails) is unsure, yet you can relegate a likelihood to every chance. This is the embodiment of the dynamic under risk. A few key ideas support this dynamic cycle:

1. Expected Value: This idea works out the normal result of choice by thinking about the probabilities of every conceivable result and its related worth (frequently financial increase or misfortune). By working out the normal worth of every choice, you can recognize the decision with the most positive typical result.
2. Risk Aversion: People are normally risk-loath. This implies we frequently focus on choices that offer an ensured, though lower, compensation over those with the potential for a higher prize yet, in addition, an opportunity of critical misfortune. Dynamic under risk thinks about this innate hazard avoidance and investigates what it means for our decisions.
3. Risk Tolerance: It is critical to note that the level of hazard resilience from one person to the other is not constant but varies. Some are more okay with getting ‘risky’ results that might not always be very good, while others prefer stability. Of equal importance is the knowledge and awareness of one’s gamble resistance needed to make rational decisions under uncertainty.

The decision theory presents decision structures, each of which endeavours to facilitate the understanding and enhancement of dynamic under risk. These systems incorporate gadgets such as choice trees, whereby various external choices and the probabilities and outcomes of the related decisions are demonstrated. Then, the expected esteem computations and the hazard resistance appraisals are incorporated into the dynamic contention.

This is why, when individuals are involved in garnering or doubt-ridden contingencies with measurable probabilities, more determined decisions can be reached if the standard for dynamic under risk is understood well. It compels people and associations to delve

into circumstances with new possibilities for the two augmentations and adversities and, in the long run, move toward more profoundly comprehended and less emotionally charged choices.

- **Challenges posed by decision-making under risk**

Managing risk and decision-making is important in many aspects of life since they are encompassing. However, they also involve numerous barriers that hinder one from making the best decision. These difficulties are the outcome of an inherent imperfection inherent to the rationing of the human mind and the nature of possibilities existing in randomness.

Another challenge relates to the accuracy of assessing probabilities. The nature of probabilities complicates their precise calculation. We often require complete information about potential solutions, and potentiality inherently has risks here. It can lead to complacency in our ability to predict events or an overestimation of possible threats. Mental inclinations also complicate morality. Bias such as mooring, that is, fixating on start data, and the tendency to seek out confirmations that support our current beliefs - the tendency to look at data in search of such an answer can distort our perception of dangers and skew our direct engagement.

The final one comes with an understanding of our reactions to risk in the nearest surroundings. Fear of risks or risk aversion, which is the tendency to try to avoid misfortunes more than risky but greater gains, can lead to over-cautious decisions. In contrast, the improbably risky discrete behaviour individuals might be drawn to choices that have big likely rewards but big drawback risks. Each of them can preclude a rational and informed assessment of alternatives.

Furthermore, the outlining of a choice can actually alter a decision in a sense. It’s in the ways that data is introduced, whether stressing likely gains or misfortunes, that various profound reactions are activated, and it is these that ultimately lead to various choices. As a result, this highlights the importance of understanding how such an outlining process can affect the context of a dynamic under risk.

Finally, the psychological consumption that stems from a desire for too many choices, referred to as choice weakness, can hinder people’s decision-making. Unfortunately, when presented with a large number of risky decisions, one’s ability to compare risks and rewards effectively can be impaired, which in turn may result in undesirable choices.

Such is the case, and considering these challenges, we are better placed to pave the way for analysing the complexity of dynamics under risk. Averting choice weariness, perceiving the mental added inclinations and deep outcomes, being familiar with the large outlining effects, and managing our feelings take vital approaches to making rational decisions even under uncertainty.

- **Knowledge Check-1**

  Fill in the blanks

  1. _____theory establishes a groundwork for settling on very much educated decisions, yet it perceives the intricacies of the human way of behaving. (**Decision**/ Frequency)
  2. These inclinations can cloud our judgment and frustrate our capacity to pursue sane decisions under_____. (Risk/ **Uncertainty**)
  3. _____like mooring (focusing on starting data) and tendency to look for predictable answers (looking for data that affirms existing convictions) can mutilate our view of dangers and slant our dynamic interaction.(**Inclinations**/ Variations)
  4. _____approach includes thinking about various conceivable future situations, even those with low probabilities.(Risk assessment/ **Situation planning**)

- **Outcome-based activity 1**

  Envision you're a business visionary going to send off another item. There's a decent likelihood of coming out on top yet, but it's also a gamble of disappointment. Make sense of the difficulties you could face in pursuing this choice under risk. Think about how mental inclinations, profound reactions, and outlining of data could impact your decision.

16.4 Decision Trees and payoff tables

- **Meaning of decision trees and payoff tables**

Decision Trees

Decision trees are an incredible asset utilised in different disciplines, going about as a visual and logical system for simply deciding. They look like flowcharts, with each inward hub addressing an inquiry or test on a particular property of the information. Branches coming from these hubs portray the potential responses, prompting ensuing

hubs or terminal hubs (leaves). These terminal hubs address the ultimate result or arrangement in view of the choices made all through the tree. By following a progression of inquiries and comparing responses to choice trees, we come to a result or forecast in a reasonable and interpretable way.

Payoff Tables

Payoff tables are dynamic apparatuses utilised in situations with various decisions and dubious results. They methodically delineate the expected outcomes of every choice option across a scope of conceivable future states. Basically, they go about as a lattice where lines address choice choices, and segments address various situations. Every cell inside the matrix shows the subsequent "result" or result (frequently financial worth, however, can address different measurements) related to picking a particular choice under a specific future state. By dissecting these adjustments, leaders can recognise the choice with the most good result, taking into account both expected rewards and dangers.

		Player 2		
		Rock	Paper	Scissors
Player 1	Rock	0 0	1 -1	-1 1
	Paper	-1 1	0 0	1 -1
	Scissors	1 -1	-1 1	0 0

- **Elements of decision tree**

Decision trees, a foundation of AI and information examination, offer an organised way to deal with characterising information and making expectations. These trees are comprised of major components that cooperate to direct dynamic interaction. Here is a breakdown of the key parts:

1. Nodes: This address choice focuses on places of division inside the tree. There are two fundamental sorts:

- Interior Hubs (Choice Nodes): These hubs house the choice models. They commonly pose an inquiry about a particular quality of the information, for example, "Is the pay more noteworthy than \$50,000?". Each inside hub branches out in numerous ways, addressing the potential responses to the inquiry.
- Terminal Hubs (Leaf Nodes): These address the information's ultimate results or characterisations. They imply the anticipated class mark (e.g., "high credit risk") or a nonstop worth (e.g., "expected house cost").

2. Branches: These rise up out of the inward hubs and address the various pathways in light of the response to the choice inquiry. Each branch prompts another hub, either one more inside hub for additional parting or a terminal hub for the ultimate result.

3. Splitting Criteria: This alludes to the standard utilised at each inner hub to separate the information. In characterisation trees, normal dividing standards include picking the property and edge that best isolates the information into particular classes. For instance, in a credit application situation, the dividing models may be to separate candidates in light of their pay being above or under a specific edge.

4. Root Node: This is the beginning stage of the tree, addressing the whole dataset and the underlying choice to be made. It basically poses the principal inquiry that sets off the ensuing expanding process.

5. Pruning: When allowed to become excessively huge, choice trees can become excessively intricate and vulnerable to overfitting. Pruning methods are utilised to decisively eliminate pointless branches, improving the tree and further developing its generalizability.

By understanding these components, you can really decipher and use choice trees for different undertakings, from client division to misrepresentation location. They offer a reasonable visual portrayal of the dynamic interaction and can be promptly figured out by both specialised and non-specialised crowds.

- **Significance of decision trees and payoff tables**

Decision trees and payoff tables are foundations of choice examination, assuming vital parts in pursuing informed decisions under vulnerability. Here is a breakdown of their importance:

Decision Trees:

1. **Representation and Clarity:** Decision trees succeed at outwardly addressing complex dynamic situations. They map out every conceivable decision, its related results, and the expected results. This reasonable visual configuration considers much more straightforward handling of the dynamic scene contrasted with crude information or extensive portrayals.
2. **Organised Analysis:** Decision trees require a certain manner of handling navigation, no matter how disorderly it may appear. This doubles because, by unambiguously laying out each decision and the expected outcomes, all relevant considerations are considered. This deliberate methodology allows for recognising not only the foreseen traps but also the unforeseen consequences.
3. **Probabilistic Analysis:** The moment a decision tree responds to the prospects, probabilities are integrated into the dynamic interaction. It is sometimes feasible to allow each branch to receive a likelihood of event, taking into account the value of expected outcomes for each choice way. This probabilistic analysis operates using the identification of the ID of a choice with the highest expected utility.
4. **Adaptability and Adaptability:** Decision trees are powerful tools that can be applied to many kinds of situations. They can be easily adjusted downwards to accrue more new data or upwards to document changes in the choice scene. This adaptability is particularly important in powerful conditions, where the condition or probability of an endeavour may enhance after a definite time.
5. **Payoff Tables:**

    1. **Quantitative Analysis:** The definitions conveyed here provide a clear quantitative system for comparing and contrasting a payoff solution with other options. They do so unambiguously reveal the anticipated winnings or losses linked to every combination of decision and outcome. This encompasses an apparent aprioristic evaluation of options based on their intrinsic utility, a measure readily derived from the resultant table.

    2. **Risk Assessment:** Payoff tables incorporate probability risks connected to every decision. This way, they help paint what is predicted to be unpleasant consequences just beside the good ones, enabling a more realistic assessment of risks and rewards.

    3**. Similar Analysis:** Payoff tables are created by comparing other choice options side by side. As each settlement was presented in a single framework,

it could be simple to distinguish the option that gives the most elevated likelihood of gain or the least expected failure, relying on the chief's risk resilience.

4. **Choice Support:** Some of the payoff tables operate like a decision aid tool in the sense that they compile all the possible effects of every decision made. To decide to banish with consequences based on certain data and do it accompanied by a clear understanding of potential risks and gains.

Decision trees and payoff tables are two wonderful tools that complement each other in the dynamic cycle. Decision trees do a good job with the first two things that an analyst might try to do with probabilities—constructing a choice scene in imagination and incorporating probabilities into an overall picture. Combined, they presented an elaborate method for tackling how best to make good and wise decisions under uncertainty.

16.5 Applications of decision theory in business strategy

Decision theory plays an important role in the development of sustainable business systems from the standpoint that it provides the framework for analysing decision-making and opting for information-driven solutions when exposed to risk. Here is a more intensive glance at its key applications: Here is a more intensive glance at its key applications:

1. **Speculation decisions:** Organisations actively evaluate venture opportunities that are prospective for the venture and range from product development, for instance, to market expansion. The choice hypothesis examines such choices by estimating the probabilities of the outcome or its converse, which is achievement or disappointment. We consider risks and possible returns, which are not always easy to compare, so methods such as expected estimations and decision trees help us focus on the ventures with the highest potential value.

2. **Risk management:** Business achievement depends on operating with flexibility in a field that is often risky. Through the use of the Choice hypothesis, organisations are invited to identify, assess and mitigate potential risks. By analysing facts and trends, one can estimate the possible and probable impact of various threats. Generally, it helps organisations classify the different risks depending on their probability of occurrence and the severity of the effects. Frameworks such as gamble captured return on capital

(RCR) link risk, which offers independent direction to ensure methodologies are concentrated on esteem creation and enduring openness to hazard.

3. **Valuing strategies:** It is crucial to set standard prices for goods and services in order to achieve optimised work output. The choice hypothesis analyses client interests, adversary appraisal, and production costs. By considering such elements and associated risks, various approaches, including earning back the original investment investigation and request gauging, are commonly used to determine expenses that bring more value and are still competitive.

4. **Promoting campaigns:** The choice hypothesis is offered a place in the reception and organisation of functional marking endeavours. Anastasia 24 An insightful portrayal of other key client socioeconomics, past mission execution analysis, and contender techniques empowers organisations to analyse different advertising channels and inform approaches. Such approaches incorporate A/B testing, which takes into account taking a gander at the possibility of various choices with the goal of enabling organisations to pick crusades with the most elevated likelihood of arriving at their coveted promoting destinations (brand character recognition, deals production).

5. **Consolidations and acquisitions (M&A):** M&A choices can be the wellsprings of mind-boggling potential benefits, yet they also have inborn risks. The choice hypothesis evaluates the extent of organisational integration, measures cooperation, and also qualitatively assesses anticipated money-making. By knowing the probable post-consolidation scenarios, these organisations can make proper decisions regarding the pursuit of the M&A and the manner in which the consolidation of the acquired organisations should be effected.

In general, decision theory consists of the fundamental independent direction tool compartment for organisations. As a result, by developing an information-driven method of treatment which again adds a certain likely identified reward and dangers linked with choice, the hypothesis of choice makes it possible for organisations to seek vulnerability as well as to look for decisions as they endeavour to receive their significant intentions and objectives.

- **Knowledge Check-2**

  State True or False

  1. Payoff tables are dynamic apparatuses utilised in situations with various decisions and dubious results. (**True**/ False)

2. Pruning methods are never utilised to decisively eliminate pointless branches, working on the tree and further developing its generalizability. (True/ **False**)
3. The choice hypothesis surveys these choices by thinking about the expected results (achievement, disappointment) and their related probabilities. (**True**/ False)
4. By taking into account the probabilities of various post-consolidation situations, organisations cannot arrive at informed conclusions about whether to seek after M&A and how to coordinate gained organisations best. (True/**False**)

- **Outcome-based activity 2**

  You are a café supervisor considering growing your menu with another vegan dish. Make a decision tree illustrating the potential results (achievement, disappointment) in light of variables like client interest and contest. Then, at that point, convert this choice tree into a payoff table that shows the benefit/misfortune related to every result for your choice to present the new dish.

16.6 Summary

- It investigates the inclinations, mental alternate ways (heuristics), and profound impacts that shape our choices, frequently driving us to go amiss from amaing sanity.
- Anticipated that worth considers probabilities and utilities should compute the typical result of a choice.
- It expects an entirely normal chief with complete information on every accessible choice, their related probabilities, and the likely results.
- We don't necessarily, in all cases, have total data, and handling complex circumstances can overpower.
- By taking into account probabilities, likely results, and individual inclinations (utilities), chiefs can move past premonitions and instinct to go with additional educated and normal decisions.
- It permits associations to assess long-haul choices, taking into account future patterns, expected difficulties, and contender activities.

- Procedures like arguing for the sake of arguing or utilising post-mortem examination (envisioning most pessimistic scenario situations) can uncover expected vulnerable sides.
- We might be enticed to stay away from decisions that convey a gamble of unfortunate results, regardless of whether the potential advantages are huge.
- Decision trees, a foundation of AI and information examination, offer an organised way to characterize information and set expectations.
- They methodically delineate the expected outcomes of every choice option across a scope of conceivable future states.
- These address the ultimate results or characterisations of the information.
- Decision trees and payoff tables are foundations of choice examination, assuming vital parts in pursuing informed decisions under vulnerability.
- Generally, decision theory gives organisations a strong tool compartment for key independent direction.

16.7 Keywords

- **Decision Theory**—Decision theory, a junction of rationale, likelihood, and human behaviour, digs into the fascinating universe of how decisions are made.
- **Utility Theory**- Utility hypothesis doles out mathematical qualities (utilities) to different results, taking into consideration an examination of choices with different outcomes.
- **Decision-making under risk—It manages circumstances in which** we face different decisions, each with known expected results, yet the probability of every result is questionable.
- **Decision Trees**- They look like flowcharts, with each inward hub addressing an inquiry or test on a particular property of the information.
- **Payoff tables**- Payoff tables are dynamic apparatuses utilised in situations with various decisions and dubious results.
- **Nodes**- These address choice focuses or places of division inside the tree.

- **Splitting Criteria**- This alludes to the standard utilised at each inner hub to separate the information.
- **Root Node**- This is the beginning stage of the tree, addressing the whole dataset and the underlying choice to be made.
- **Pruning**- Decision trees can turn out to be excessively intricate and defenceless to overfitting whenever permitted to become overly huge.

16.8 Self-Assessment questions

1. What do you mean by decision theory?
2. What are the steps taken to reduce errors by decision-making under uncertainty?
3. What challenges does decision-making under risk pose?
4. What is the significance of decision trees and payoff tables?
5. What are the major applications of decision theory in business strategy?

16.9 References/ Reference Reading

- Sharma, J. K. *Operations Research: Theory and Applications*. 6th ed., Macmillan Publishers India, 2017.
- Taha, Hamdy A. *Operations Research: An Introduction*. 10th ed., Pearson Education India, 2017.
- Kothari, C. R., and Gaurav Garg. *Research Methodology: Methods and Techniques*. 4th ed., New Age International Publishers, 2019.
- Hillier, Frederick S., and Gerald J. Lieberman. *Introduction to Operations Research*. 11th ed., McGraw Hill Education, 2020.
- Vohra, N. D. *Quantitative Techniques in Management*. 5th ed., Tata McGraw Hill Education, 2017.