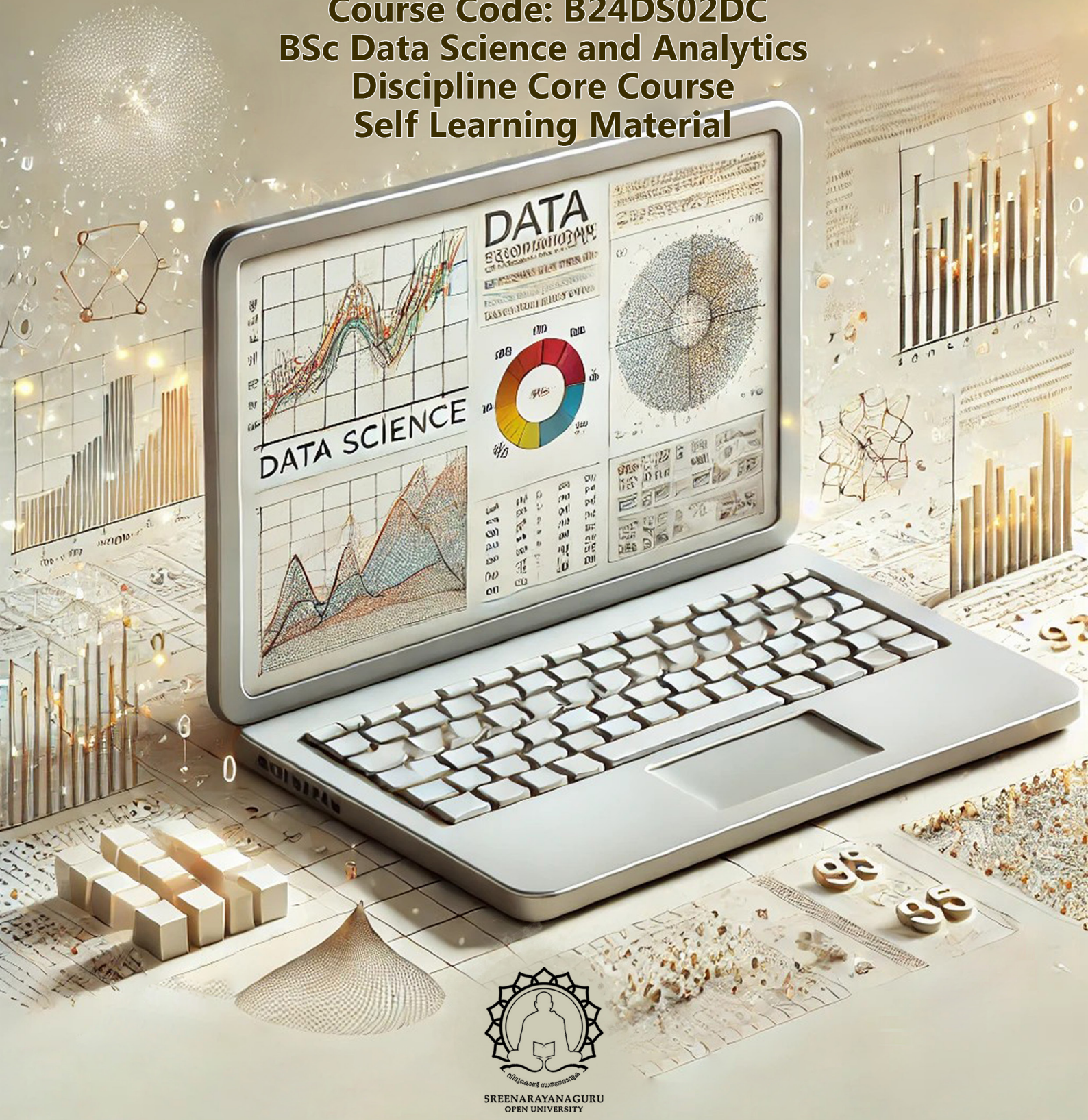
Computational Foundations for Data Science

Course Code: B24DS02DC
BSc Data Science and Analytics
Discipline Core Course
Self Learning Material



SREENARAYANAGURU
OPEN UNIVERSITY

SREENARAYANAGURU OPEN UNIVERSITY

The State University for Education, Training and Research in Blended Format, Kerala

SREENARAYANAGURU OPEN UNIVERSITY

Vision

To increase access of potential learners of all categories to higher education, research and training, and ensure equity through delivery of high quality processes and outcomes fostering inclusive educational empowerment for social advancement.

Mission

To be benchmarked as a model for conservation and dissemination of knowledge and skill on blended and virtual mode in education, training and research for normal, continuing, and adult learners.

Pathway

Access and Quality define Equity.

Computational Foundations for Data Science

Course Code: B24DS02DC

Semester - I

Discipline Core Course
Undergraduate Programme
BSc Data Science and Analytics
Self Learning Material

SREENARAYANAGURU
OPEN UNIVERSITY

SREENARAYANAGURU OPEN UNIVERSITY

The State University for Education, Training and Research in Blended Format, Kerala

MESSAGE FROM VICE CHANCELLOR

Dear learner,

I extend my heartfelt greetings and profound enthusiasm as I warmly welcome you to Sreenarayanaguru Open University. Established in September 2020 as a state-led endeavour to promote higher education through open and distance learning modes, our institution was shaped by the guiding principle that access and quality are the cornerstones of equity. We have firmly resolved to uphold the highest standards of education, setting the benchmark and charting the course.

The courses offered by the Sreenarayanaguru Open University aim to strike a quality balance, ensuring students are equipped for both personal growth and professional excellence. The University embraces the widely acclaimed "blended format," a practical framework that harmoniously integrates Self-Learning Materials, Classroom Counseling, and Virtual modes, fostering a dynamic and enriching experience for both learners and instructors.

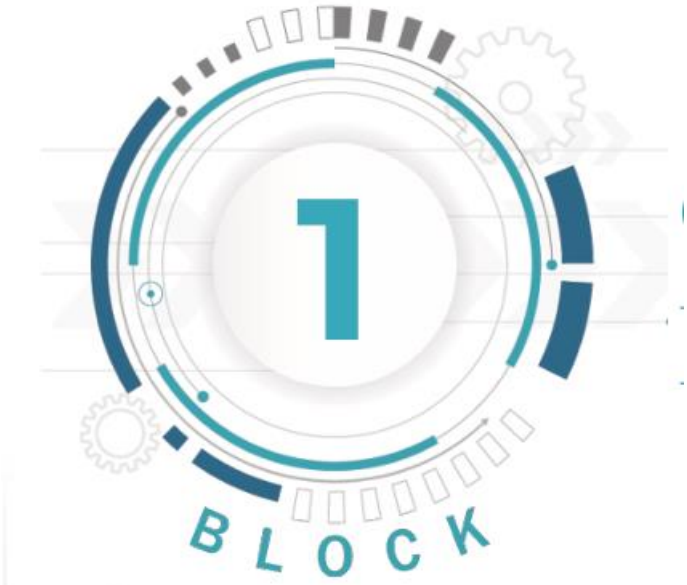
The University is committed to providing an engaging and dynamic educational environment that encourages active learning. The Study and Learning Material (SLM) is specifically designed to offer you a comprehensive and integrated learning experience, fostering a strong interest in exploring advancements in information technology (IT). The curriculum has been carefully structured to ensure a logical progression of topics, allowing you to develop a clear understanding of the evolution of the discipline. It is thoughtfully curated to equip you with the knowledge and skills to navigate current trends in IT, while fostering critical thinking and analytical capabilities.The Self-Learning Material has been meticulously crafted, incorporating relevant examples to facilitate better comprehension.

Rest assured, the university's student support services will be at your disposal throughout your academic journey, readily available to address any concerns or grievances you may encounter. We encourage you to reach out to us freely regarding any matter about your academic programme. It is our sincere wish that you achieve the utmost success.

Regards,
Dr. Jagathy Raj V. P.

01-01-2025

Contents



Introduction to Statistics

UNIT 1

Definition and Importance of Statistics

Learning outcome

The student will be able:

- Recall the historical origins of statistics.
- Define key statistical terms (population, sample, variable).
- Identify descriptive and inferential statistics.
- List main roles of statistics in data analysis.
- Recognize examples of statistics in various fields.

Pre-requisites

Imagine you're organizing a large event, like a school fair, where you want everything to be well-prepared. You'll need to estimate how many people will attend, plan how much food is needed, and decide on the number of games or activity stations. But how can you make these decisions without knowing the exact numbers? This is where a basic understanding of statistics becomes helpful. Just as a weather forecast helps us decide what to wear, statistics gives us tools to understand patterns and make informed choices. You've likely already used similar skills when tracking your own budget, calculating grades, or managing your time. These habits make daily life easier by turning numbers into useful information. In the same way, statistics can help people in many fields make sense of large amounts of data. This unit will show how these methods help answer important questions across areas like science, business, and everyday life. Let's explore how statistics shapes our understanding!

Key Concepts

Statistics, Data Collection, Patterns, Prediction, Decision Making, Estimation

Discussion

1.1.1 Overview of Statistics

The origins of statistics are deeply rooted in the practical needs of ancient civilizations for data collection and record-keeping. Data for administrative purposes have been collected as far back in history at least to ancient Egypt, and there is fairly well recorded information from Babylonia and China whose census taking was associated with resource allocation, labour management or taxation. Both Greeks and Romans had some forms of censuses with the Romans having periodic census for military recruitment as well as tax purposes. Islamic scholars developed survey sampling and data collection patterns during the middle ages, while European states held systematic records for governance. The 17th century saw the origin of modern statistics with John Graunt's analysis of mortality data and William Petty's application of statistical methods to economics. Pierre-Simon Laplace and Carl Friedrich Gauss contributed to probability theory and statistical methods in the 18th century. It wasn't until the 19th century that things began to get a little more sophisticated, such as Adolphe Quetelet applying statistics in the social sciences and Francis Galton giving us correlation and eventually regression. It was in the 20th century that statistics were formalized as a separate scientific field, with works by Ronald A. Fisher, Jerzy Neyman and Egon Pearson on modern inferential statistics. These days statistics is found everywhere and also with progress in other fields like big data or machine learning have made it a tool of importance for evidence-based decision making across the industry.

1.1.2 Definition of Statistics

Statistics is a branch of Mathematics that falls into collection, analysis, interpretation, presentation and organizing data. Basically, data science is the study of extracting information from large amounts of unstructured or structured data.

Descriptive Definition: Statistics is the science of collecting, describing, and interpreting data to summarise and communicate information.
Inferential Definition: Statistics involves using data from a sample to draw conclusions about a larger population, utilising probability theory to estimate parameters and test hypotheses.

1.1.3 The Role and Application of Statistics

Summarizing Data: Statistics plays a crucial role in organizing large sets of data into meaningful summaries. Through descriptive measures like mean, median, and mode, statistics condenses complex data into understandable formats, allowing us to see key insights at a glance. This is essential for transforming raw data into usable information that can be easily communicated and interpreted.

Analyzing Data: With statistical tools, data analysis becomes a structured process that helps identify patterns and relationships within datasets. By applying statistical techniques, analysts can uncover trends, correlations, and causal relationships that might otherwise go unnoticed. This analysis is foundational for research, enabling scientists and professionals to understand

complex behaviors and phenomena.

Making Inferences: Inferential statistics allows us to draw conclusions about an entire population based on a smaller sample of data. By applying probability-based methods, we can estimate population parameters and test hypotheses, even with limited data. This capability makes statistics essential for research and policy-making, where making informed predictions or generalizations is often required.

Decision Making: In uncertain situations, statistics provides a reliable basis for making informed decisions. Statistical methods quantify uncertainty and offer evidence-based insights that guide choices in areas such as business, healthcare, and public policy. This enables decision-makers to assess risks, predict outcomes, and select actions that align with their goals and available information.

1.1.4 Key Terms in Statistics

Population: In statistics, a population refers to the complete set of individuals or observations that are of interest in a particular study. It encompasses all members of a defined group, which may be large or small, depending on the research focus. If a study aims to understand the academic performance of university students, the population would include every student enrolled in that institution. Understanding the population is crucial because it sets the foundation for sampling and statistical analysis. By defining the population clearly, researchers can ensure that their findings are relevant and applicable to the group they intend to study.

Sample: A sample is a smaller, manageable subset of the population selected for analysis in a statistical study. It is crucial for researchers to gather data from samples because studying an entire population can be impractical or impossible. For example, if a university wants to survey student opinions on campus facilities, they might select a sample of 200 students from the entire student body. The key to effective sampling lies in ensuring that the sample accurately represents the population, which can be achieved through random sampling or other methods. By analyzing a sample, researchers can draw conclusions about the broader population without needing to collect data from everyone.

Variable: A variable is any characteristic or attribute that can take on different values in a study. Variables are essential for statistical analysis as they provide the data points that researchers seek to measure and analyze. Examples of variables include age, height, income, and education level, among many others. Variables can be classified into different types, such as quantitative variables (which represent numerical values) and qualitative variables (which represent categories or qualities). Understanding the nature of variables helps researchers determine the appropriate statistical methods to use in their analyses, as different types of variables require different approaches.

Data: Data refers to the actual values or observations collected from variables during a study. It serves as the foundation for statistical analysis and can be classified into two primary types: quantitative and qualitative. Quantitative data includes numerical values that can be measured and compared, such as test scores or temperature readings. On the other hand, qualitative data consists of categorical values that describe characteristics or attributes, such as favorite colors or types of cuisine. The accuracy and reliability of data collection are vital, as the quality of the data directly impacts the validity of the conclusions drawn from the analysis.

1.1.5 Types of Statistics

1.1.5.1 Descriptive Statistics

Descriptive statistics is a part of statistics that focuses on summarizing and presenting data clearly. It helps us understand the main features of a dataset by using different methods. One important method is measuring central tendency, which includes the mean (average), median (middle value), and mode (most common value). Another method is measuring variability, which tells us how spread out the data points are, using range, variance, and standard deviation. We also use graphical representations, like histograms and pie charts, to show data visually. Descriptive statistics are essential because they provide a clear overview of the data, making it easier to interpret.

1.1.5.2 Inferential Statistics

Inferential statistics allows researchers to make predictions or conclusions about a larger group based on data collected from a smaller sample. This branch uses probability theory to help us understand how we can generalize findings beyond the immediate data. Key techniques include estimation, which helps us guess about population parameters using point estimates and confidence intervals. Another important technique is hypothesis testing, where we check assumptions about a population to see if there is enough evidence to support or reject a claim. Common tests in hypothesis testing include t-tests, which compare averages, and chi-square tests, which look at relationships between categories. By using inferential statistics, researchers can make informed decisions and validate their findings, making it a vital part of scientific research.

1.1.6 Importance of Statistics

Statistics is a unique subject that is closely connected to human life and development. It provides essential tools for understanding the world around us and helps us make sense of complex information. By using statistics, we can improve business processes, study technological advancements, and gain insights into social issues and environmental challenges. The methods of statistics allow us to collect and organize data systematically, which is crucial for making informed decisions. Overall, statistics is not just a collection of numbers; it is a way to interpret and understand the many aspects of our lives.

The importance of statistics cannot be overstated. It has become an essential part of modern life and contributes to progress in various fields. Statistics provides the necessary techniques for gathering, organizing, and analyzing data, which are vital for understanding numerous phenomena and processes. From guiding business strategies to informing scientific research, statistics plays a critical role in many areas. Additionally, it helps policymakers make evidence-based decisions that affect society and the environment. In summary, statistics is important in many ways, impacting our understanding of the world and improving our ability to address challenges. Few important aspects of Statistics is listed below:

1.1.7 Decision Making

1.1.7.1 Informed Choices

Statistics is an important tool for making informed decisions in different areas, such as business and public policy. By analyzing data, decision-makers can look at the possible outcomes of

various actions and choose the best option. This approach helps organizations rely on facts instead of guesses when making choices. For example, in business, statistical analysis can help companies decide which products to launch by looking at market research and consumer behavior data. This method allows businesses to better meet customer needs and improve their chances of success.

1.1.7.1 Risk Assessment

Statistics also helps in measuring and managing risks. By understanding how likely different events are and what impact they might have, organizations can create plans to reduce negative effects. This is especially important in finance, where risk assessment models are used to manage investment portfolios and predict changes in the market. Using statistical methods, companies can find potential risks and make smart decisions to protect their investments. Financial analysts use statistical data to forecast market trends, helping them adjust their strategies to avoid losses. Overall, using statistics for risk assessment helps organizations handle uncertainty and make better financial choices.

1.1.8 Understanding Variability

Quantifying Uncertainty: Statistics helps in understanding and quantifying the variability and uncertainty inherent in data. This is crucial for making accurate predictions and inferences. Example: In weather forecasting, statistical models account for various uncertainties to predict future conditions with a certain level of confidence.

Identifying Patterns: Statistical techniques are used to identify patterns and trends within data, which can lead to new insights and discoveries. Example: In healthcare, analysing patient data can reveal trends in disease outbreaks or the effectiveness of treatments.

1.1.9 Quality Control

Improving Processes: Quality control is an essential aspect of production and manufacturing that focuses on maintaining and improving product quality. Statistics play a critical role in this process by providing tools to monitor, evaluate, and refine production activities. Statistical methods, such as control charts and process capability analysis, are commonly used to ensure that products consistently meet established quality standards. Control charts help track process stability over time, indicating any deviations that could lead to defects or inconsistencies. Meanwhile, process capability analysis assesses whether a process can operate within specified limits, enhancing reliability and efficiency in production.

> An example of these methods in action is Six Sigma, a widely adopted approach in manufacturing. Six Sigma employs statistical tools to identify and reduce defects, ultimately improving product quality and customer satisfaction.

Standardisation: Standardisation is another key area where statistical analysis supports quality control. By establishing consistency in processes and products, standardisation promotes reliability and quality across batches. In the pharmaceutical industry, statistical methods ensure that each batch of medication meets strict safety and efficacy requirements. This practice guarantees that medications are safe, effective, and consistent, which is crucial for patient health and regulatory compliance. Through these applications, statistics underpin quality control efforts, fostering trust and dependability in various industries.

1.1.10 Research and Development

Designing Experiments: Research and development heavily rely on statistics, particularly in the design and execution of experiments and clinical trials. Statistics is fundamental to planning experiments, determining the optimal sample sizes, randomising treatments, and analysing results to ensure valid and reliable conclusions. Well-designed experiments are essential for making informed decisions and minimizing bias in research findings.

A common example of this in practice is found in medical research, where randomized controlled trials (RCTs) are used to assess the efficacy of new treatments. In RCTs, statistical methods ensure that treatments are assigned randomly to minimize bias and accurately reflect treatment effects. This approach provides a robust framework for evaluating medical interventions, allowing researchers to identify treatments that are truly beneficial and safe. Through these applications, statistics contribute significantly to advancing research by ensuring experimental rigor and credibility.

Data Analysis: Statistical techniques are vital in analyzing research data, testing hypotheses, and validating scientific theories across various fields. By applying statistical methods, researchers can interpret data accurately, identify patterns, and evaluate the strength of their findings. Hypothesis testing, allows researchers to assess whether their results are statistically significant and not due to chance. This process is essential for making informed conclusions and advancing scientific knowledge.

In psychology, for example, statistical analysis is commonly used to study behavior patterns and validate theories of human cognition. Researchers might analyze data from experiments or surveys to determine if observed behaviors align with theoretical predictions about cognition and behavior. By rigorously applying statistical techniques, psychology and other fields can develop, refine, and support theories, leading to a deeper understanding of complex phenomena.

1.1.11 Policy Making

Evidence-Based Policy: Governments and organizations depend heavily on statistical data to create and evaluate policies that address societal needs effectively. Statistical analysis provides a foundation of empirical evidence, enabling policymakers to make informed decisions that are responsive to real-world conditions. By analyzing data, governments can understand population trends, economic shifts, and social challenges, all of which guide the development of policies that are not only effective but also sustainable.

For example, census data play a crucial role in shaping public policies related to infrastructure, education, and healthcare. By examining detailed population data, policymakers can determine where to build schools, allocate healthcare resources, and develop transportation networks to serve communities efficiently. This reliance on statistical analysis ensures that policy decisions are grounded in reliable data, promoting policies that are tailored to meet the diverse needs of society.

Resource Allocation: Statistics is essential in the efficient allocation of resources, as it helps identify areas of need and evaluate the effectiveness of various interventions. By analyzing data, organizations and governments can prioritize resources strategically, ensuring they are directed toward areas with the greatest impact. Statistical methods allow decision-makers to forecast demands, assess intervention outcomes, and adjust resources accordingly, promoting a more efficient and responsive allocation process.

In public health, for example, statistical models play a crucial role in vaccine distribution during an outbreak. These models analyze factors such as population density, infection rates, and risk

groups to determine where vaccines are most urgently needed. This targeted approach ensures that limited resources, like vaccines, are used effectively to control the spread of disease and protect vulnerable populations. Through such applications, statistics enables resource allocation that maximizes impact and improves overall outcomes.

1.1.12 Economic Planning

Forecasting: Statistics plays a crucial role in forecasting economic trends, including GDP growth, inflation rates, and unemployment levels. Accurate forecasts provide valuable insights into the future state of the economy, which are essential for informed economic planning and policy-making. By analyzing historical data and identifying patterns, statistical models enable economists and policymakers to anticipate changes and prepare appropriate responses to potential challenges or opportunities.

For example, central banks rely on statistical models to predict economic conditions, allowing them to adjust monetary policies proactively. By forecasting inflation trends or shifts in employment rates, central banks can make decisions on interest rates, currency supply, and other factors to stabilize the economy. Through such applications, statistics empowers economic institutions to guide policy decisions that foster stability and growth.

Market Analysis: Businesses rely on statistical methods to analyze market trends, understand consumer preferences, and assess competitive dynamics. By gathering and interpreting data on these factors, companies can make informed strategic decisions that position them to capture market opportunities effectively. Statistical analysis helps businesses identify patterns, forecast demand, and respond to shifts in the market, enabling them to stay competitive and meet consumer needs efficiently.

Retail companies analyze sales data to detect emerging consumer trends, allowing them to adjust inventory levels accordingly. By using statistical insights, retailers can ensure popular products are available while minimizing excess stock, leading to optimized inventory management and better customer satisfaction. This data-driven approach helps businesses adapt quickly to market changes, maximizing profitability and strengthening their market presence.

1.1.13 Social Sciences

Understanding Social Phenomena: Statistics is foundational in the social sciences, providing essential tools to study human behavior, social trends, and societal issues. By employing statistical methods, social scientists can collect, organize, and analyze data to gain insights into various aspects of society. This data-driven approach enables researchers to explore complex social phenomena, identify trends, and draw meaningful conclusions about human interactions and societal dynamics. For example, sociologists use statistical surveys to study demographic changes and social attitudes. These surveys collect data on aspects such as age, education, income, and cultural beliefs, offering a comprehensive view of social patterns and shifts. Through statistical analysis, sociologists can uncover insights into how societies evolve and how attitudes change over time, providing a valuable basis for understanding and addressing societal challenges.

Policy Impact Assessment: Statistical analysis plays an important role in evaluating the impact of social policies and programs, ensuring that they achieve their intended outcomes. By applying statistical methods, researchers can assess whether specific interventions are effective in addressing societal needs and improving conditions for targeted populations. This evaluation process involves collecting data before and after the implementation of policies or programs,

allowing for a comparison that highlights their effectiveness. Educational researchers often use statistical methods to assess the effectiveness of new teaching methods and curricula. By analyzing student performance data, attendance rates, and engagement levels, researchers can determine whether these innovations lead to improved educational outcomes. This empirical evidence informs educators and policymakers about the success of specific initiatives and guides future decisions and resource allocation to enhance educational practices. Through rigorous statistical evaluation, the impact of social policies and programs can be understood and refined, promoting continuous improvement in societal initiatives.

1.1.14 Environmental Science

Climate Modelling: Statistical techniques are employed to model and predict climate change, assisting scientists in understanding long-term trends and the impacts of global warming. By applying these methods, researchers can analyze complex environmental data to identify patterns and correlations that inform their understanding of climate dynamics. This analytical approach is vital for forecasting future climate conditions and evaluating potential effects on ecosystems and human societies. For example, environmental statisticians analyze temperature data to predict future climate patterns and their possible impacts. By examining historical temperature records and incorporating various factors such as greenhouse gas emissions and natural variability, they can create models that simulate future scenarios. This information is crucial for policymakers and stakeholders, as it helps them make informed decisions regarding climate adaptation and mitigation strategies. Through the application of statistical techniques, scientists can better understand climate change and its implications, supporting efforts to address this global challenge.

Resource Management: Statistics plays an important role in managing natural resources by analyzing data related to resource availability, usage, and sustainability. By employing statistical methods, researchers and resource managers can gather insights that guide effective management practices. This analytical approach enables the evaluation of current resource conditions and helps forecast future trends, ensuring that resources are used responsibly and sustainably. In the fishing industry, statistical models are used to estimate fish populations and establish sustainable catch limits. By analyzing data on fish stock assessments, breeding patterns, and environmental factors, fisheries can determine the health of fish populations and set quotas that prevent overfishing. This approach not only helps maintain ecological balance but also supports the livelihoods of communities reliant on fishing. Through the application of statistics, effective management of natural resources can be achieved, promoting sustainability and long-term viability.

1.1.15 Practical Applications

Business and Economics: In the fields of business and economics, statistics is extensively used for various purposes, including market research, quality control, financial analysis, and economic forecasting. In market research, statistical methods help businesses understand consumer behavior, preferences, and trends. By analyzing survey data and sales figures, companies can make informed decisions about product development, marketing strategies, and target demographics.

Quality control relies on statistical techniques to monitor and improve production processes. By applying methods such as control charts and process capability analysis, businesses can ensure that products meet quality standards and identify areas for improvement. In financial analysis, statistics aids in evaluating investment opportunities, assessing risk, and analyzing market trends.

Financial analysts use statistical tools to interpret data and make forecasts about future performance, guiding strategic investment decisions.

Economic forecasting utilizes statistical models to predict future economic conditions, such as GDP growth, inflation rates, and unemployment levels. These forecasts are essential for policymakers and businesses alike, as they provide valuable insights for planning and decision-making.

Consumer Behaviour Analysis:

Surveys and Questionnaires: Consumer behaviour analysis is a vital aspect of marketing and product development, relying heavily on statistical methods to gather and interpret data about consumer preferences and habits. Companies often design surveys and questionnaires to collect valuable insights from their target audience. By employing statistical techniques to analyse responses, businesses can better understand consumer preferences, buying habits, and satisfaction levels. This information is crucial for tailoring products and services to align with market demands, ensuring that offerings resonate with consumers.

Focus Groups: In addition to surveys, focus groups provide another layer of understanding consumer behavior. Statistical analysis of focus group data helps businesses gain insights into consumer attitudes and perceptions. By examining the discussions and feedback from focus group participants, companies can identify trends, preferences, and potential areas for improvement in their products or marketing strategies. This qualitative data, when analysed statistically, enhances the overall understanding of consumer behavior and informs effective product development and marketing efforts. Through these approaches, statistical methods empower businesses to make informed decisions that cater to consumer needs and drive market success.

Market Segmentation: Market segmentation is a key strategy for businesses seeking to understand and effectively target their audience. Statistical techniques such as cluster analysis and factor analysis play significant roles in this process.

Cluster Analysis is a powerful statistical method that helps businesses segment their market into distinct groups based on various criteria, including demographic, psychographic, and behavioral factors. By grouping consumers with similar characteristics, companies can tailor their marketing campaigns to resonate with specific segments. This targeted approach enhances the effectiveness of marketing efforts, ensuring that messaging and products align closely with the needs and preferences of each segment.

Factor Analysis: Factor analysis complements cluster analysis by identifying underlying factors that influence consumer behavior. By analyzing data on various consumer attributes, businesses can uncover the key drivers behind purchasing decisions and preferences. This insight allows companies to develop more focused marketing strategies that address the specific motivations and interests of their target audience.

Statistics is essential across various fields, playing a vital role in healthcare, social sciences, environmental science, and engineering.

In healthcare, statistics is fundamental for clinical trials, epidemiological studies, and healthcare policy planning. Clinical trials use statistical methods to design experiments that assess the safety and efficacy of new treatments, ensuring that results are reliable and valid. Epidemiology relies on statistical analysis to track disease patterns, identify risk factors, and evaluate the

effectiveness of public health interventions. Furthermore, statistical insights are crucial for healthcare policy planning, helping to allocate resources effectively and improve health outcomes.

In the social sciences, statistics is utilized in disciplines such as psychology, sociology, and education to analyse human behaviour and social trends. Researchers employ statistical techniques to collect and interpret data, allowing them to identify patterns and relationships within populations. This analysis helps inform theories and practices, enhancing our understanding of societal issues and human interactions.

In environmental science, statistics is applied in climate studies, pollution analysis, and resource management. Statistical models enable scientists to analyse environmental data, track changes in climate patterns, and assess the impact of human activities on ecosystems. This information is essential for developing strategies to mitigate environmental issues and promote sustainable practices.

In engineering, statistics is important for quality assurance, reliability testing, and process optimization. Statistical methods help engineers assess the performance and reliability of products, ensuring they meet safety and quality standards. By analysing data from tests and experiments, engineers can optimize processes, improve efficiency, and reduce costs.

Recap

- **Definitions of Statistics:**
  - **Descriptive Definition:** The science of collecting, describing, and interpreting data.
  - **Inferential Definition:** Using sample data to draw conclusions about a larger population through probability theory.
- **The Role of Statistics:**
  - **Summarizing Data:** Helps condense large data sets into understandable formats.
  - **Analyzing Data:** Identifies patterns and relationships within data.
  - **Making Inferences:** Enables conclusions about populations based on sample data.
  - **Decision Making:** Aids informed decision-making under uncertainty.
- **Key Terms in Statistics:**
  - **Population:** Entire set of individuals or observations of interest (e.g., all students in a university).
  - **Sample:** Subset of the population selected for analysis (e.g., 200 students from the university).
  - **Variable:** Any characteristic that can take different values (e.g., age, height).
  - **Data:** Values or observations collected from variables; can be quantitative or qualitative.
- **Types of Statistics:**

- **Descriptive Statistics:** Focuses on summarizing data (measures of central tendency and variability).
- **Inferential Statistics:** Deals with making predictions about a population based on sample data (estimation and hypothesis testing).

➢ **Importance of Statistics:**

- **Decision Making:** Supports informed choices in business and public policy.
- **Understanding Variability:** Quantifies uncertainty and identifies patterns in data.
- **Quality Control:** Monitors and improves production processes.
- **Research and Development:** Essential in designing experiments and analyzing research data.
- **Policy Making:** Informs evidence-based policy formulation and resource allocation.
- **Economic Planning:** Used for forecasting economic trends and market analysis.
- **Social Sciences:** Studies human behavior and evaluates social policies.
- **Environmental Science:** Models climate change and manages natural resources.
- **Practical Applications:** In business, healthcare, social sciences, environmental science, and engineering.

➢ **Consumer Behaviour Analysis:**

- **Surveys and Questionnaires:** Collect and analyze responses to tailor products and services.
- **Focus Groups:** Provide insights into consumer attitudes for product development.
- **Market Segmentation:** Uses cluster and factor analysis for targeted marketing strategies.

Objective Type Questions

1. Which ancient civilization is known for the earliest recorded data collection for administrative purposes?
2. Who is considered the pioneer of modern statistics due to his analysis of mortality data?
3. Define statistics in terms of its primary functions.
4. What is the primary difference between descriptive and inferential statistics?
5. Name one statistical method used to measure central tendency.
6. What is the purpose of hypothesis testing in inferential statistics?
7. Give an example of a qualitative variable in statistics.
8. Why is sampling important in statistical studies?
9. How does descriptive statistics help in understanding data?
10. Mention one application of statistics in real-world decision-making.
11. Define informed decision-making in the context of statistics.
12. How does statistics help in risk assessment?
13. What role does statistics play in identifying patterns within data?
14. Explain the importance of control charts in quality control.
15. How do randomized controlled trials (RCTs) use statistics in medical research?
16. Why is statistical analysis important in policy-making?
17. How does statistical forecasting assist in economic planning?
18. Explain how statistics supports climate modeling in environmental science.
19. Describe the use of statistical methods in consumer behavior analysis.
20. What is the significance of statistical methods in resource management?

Answers

1. Ancient Egypt
2. John Graunt
3. Statistics involves the collection, analysis, interpretation, presentation, and organization of data.
4. Descriptive statistics summarize data, while inferential statistics use samples to draw conclusions about a population.
5. Mean (average)
6. To test assumptions about a population and determine if there is enough evidence to support or reject a claim.
7. Favorite color
8. It allows researchers to study a smaller, representative group instead of the entire population, making data collection practical.
9. It summarizes large datasets using measures like mean, median, mode, and graphical representations.
10. Business decision-making based on market trends.
11. Using data to make objective and logical decisions.
12. By analyzing historical data to estimate potential risks.
13. It identifies trends, correlations, and anomalies in data.
14. It monitors process stability and detects variations.
15. By analyzing treatment effects and comparing groups.
16. It provides data-driven insights for policy formulation.
17. By predicting economic trends using past data.
18. It helps model climate changes using statistical techniques.
19. It analyzes purchasing patterns and preferences.
20. It ensures optimal use of available resources.

Suggested Readings

1. Newbold, Paul, William L. Karave, and Betty Thorne. *Statistics for Business and Economics*. 9th ed., Pearson, 2013.
2. Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed., Springer, 2009.
3. Freedman, David A., Robert Pisani, and Roger Purves. *Statistics*. 4th ed., W.W. Norton & Company, 2007.
4. Wheelan, Charles. *Naked Statistics: Stripping the Dread from the Data*. W.W. Norton & Company, 2013.
5. Bruce, Peter, and Andrew Bruce. *Practical Statistics for Data Scientists: 50 Essential Concepts*. O'Reilly Media, 2020.

UNIT-2

Methods of Data Collection and Sampling

Learning outcome

The student will be able:

- **Identify** different methods of data collection used in research.
- **List** various types of surveys and their characteristics.
- **Define** what constitutes an experiment and its key components.
- **Recognize** the types of observational studies commonly used in research.
- **Describe** different sampling methods employed in data collection.
- **Name** the advantages and disadvantages of specific data collection methods.

Pre-requisites

Imagine you're a treasure hunter setting out to survey a vast forest. You know it's impossible to explore every tree, but you have a few tricks up your sleeve to make your task easier. Would you randomly explore specific areas, or follow a trail to cover consistent intervals? Sampling techniques work in much the same way, helping researchers "explore" a population by focusing on parts of it without needing to observe everyone.

In **Cluster Sampling**, it's like selecting entire groves within the forest—surveying each tree in chosen groves rather than random individual trees across the entire area. Imagine you pick a few districts in a state and then study every household within them. This approach saves time and resources, though it works best when the chosen clusters represent the whole population fairly well.

With **Systematic Sampling**, it's as if you're traveling along a straight path in the forest, stopping every nth tree to inspect. Suppose you lined up your class alphabetically and picked every 5th student, you'd be doing something similar: using a consistent pattern that can be both efficient and simple, provided your list doesn't have hidden patterns.

Each method has its strengths and best uses, much like a hunter choosing between strategies to cover a vast forest efficiently. By knowing these techniques, researchers can pick the approach that best fits their needs, balancing precision, efficiency, and practicality to gather meaningful insights.

Keywords:

Experiments, Surveys, Simple Random Sampling, Stratified Sampling, Cluster Sampling, Systematic Sampling, Population, Sample, Sampling Error,

Discussion

Data is always needed for any type of work. But, in order to obtain conclusive data, it needs to be strong and reliable. There are three main ways of collecting data: surveys, experiments, and observational studies.

A survey is a method, used to gather information by asking people a series of pre-set questions. There are several types of surveys, conducted over a telephone, online or computer interviews, in-house by the interviewer, or using focus groups.

Telephone surveys let researchers contact people directly, giving a personal touch while needing fewer interviewers, which helps reduce costs. Some systems even have special software to automatically generate reports, making the process more efficient. However, as more people use mobile phones—especially those that can't be reached through landline-based surveys—telephone surveys face limitations in reaching a broad, diverse population. This shift makes it harder to get a complete view of the population's opinions or behaviors.

Online or computer interview surveys provide automatic recording of the answers which cut costs, though the software is also costly, as well as the opportunity to generate automatic reports. In-house interviews, either self-administered by filling in a questionnaire or administered by the interviewer, mean cheaper costs but either limited choice of people or limited representativeness.

1.2.1 Methods of Data Collection

Data is always needed for any type of work. But, in order to obtain conclusive data, it needs to be strong and reliable. There are three main ways of collecting data:

- Surveys
- Experiments
- Observational Studies

1.2.2 Surveys

A survey is a method that is used to collect information by asking people a set of questions formulated in advance. There are several types of surveys: via telephone, online or computer interview, in-home interview with an interviewer, or using respondents' focus groups. A survey is a method of inquiring a group of people that is a sample of a population about some questions concerning their opinions, experiences, or behaviour. A survey is a flexible tool used to collect information on various topics. Surveys are the foundation of social research, therefore, today there are numerous options that can be used. That is why the choice of the type of survey is, first of all, instrumental in ensuring that all information one can receive is correct and can be generalized.

1.2.2.1 Types of Surveys

1. **Telephone Surveys:** Telephone surveys offer respondents personal contact, reduced number of interviewers, which in its turn often means lower costs, and some are equipped with software that helps in generating reports. However, the increased numbers of mobile phones, particularly the cell-phones which cannot be used for telephone surveys do not provide a comprehensive population scope.

2. **Face-to-Face Surveys:** In-person surveys are those interacting directly with the survey takers. They allow researchers to clarify challenging questions and ask for detailed responses. At the same time, this is the most labour- as well as time-consuming and thus costly way to conduct a survey. In answer, they also demonstrate the highest response rates, since the survey taker is in personal contact with the respondent. This feature enables presentation of visual aids and other kinds of support, detailed explanations, and, where necessary, extra support to clarify the question and the answer, simultaneously collecting the information from the answer as well as non-verbal information provided by the respondent's body. Overall, simultaneously with the telephone survey, it is relatively challenging in terms of time and money.

3. **Online Surveys:** Distributed electronically, enabling the fast and cost-effective data collection and potentially allowing for a greater reach, however, with the potential downsides such as lower response rates and heightened risk of bias. Online surveys are the type of survey allowing for relatively quick and budget-friendly data collection from a relatively large sample. This method may be used to study samples that are geographically separated since it is easy to target the remote host at the same time. The feature of greater anonymity due to lack of social presence of the interviewer and the respondent should also be mentioned, along the electronic nature of the survey, which allows for rapid advances in data collection and analysis. However, there are some downsides to the method: as a rule of thumb, online surveys tend to have lower response rates compared to face-to-face interviews and telephone surveys. Second, samples may be biased if the population of interest does not have access to the technology, the Internet, or both. And last, the distraction by other activities or people may influence the quality of survey completion.

4. **Other Survey Methods:** Apart from the mentioned above methods for surveying the population, the researchers also have a toolbox of additional survey options that can be used in specific circumstances.
   First, mail surveys have been one of the most traditional survey methods since the last century. However, as the tendency of transitioning to digital devices persists, the rates of mail surveys declined, and the responses were received only from representatives of the older population.

   Second, mobile phone usability brought an opportunity for quick and more focused message distribution (SMS). SMS surveys that are delivered on a mobile are becoming more and more popular.

   Finally, kiosk surveys are another convenient option for gathering further input in public places. Many organizations and businesses use kiosks for conducting surveys, and due to their limited mobility and reach, kiosk responses come from the group of those who visit the place.

1.2.3 Survey Formats

1.2.3.1 Structured Surveys

Closed-ended questions are designed with predetermined answer choices, enabling straightforward data analysis and quantification. Common examples include multiple-choice questions and Likert scale items, where respondents can express their level of agreement from "strongly agree" to "strongly disagree." One of the primary advantages of using closed-ended questions is that they produce standardized responses, which facilitate comparison and statistical analysis across different participants. Additionally, these questions help minimize interviewer bias and reduce errors during data collection.

However, a significant disadvantage is that closed-ended questions limit the richness and depth of responses that open-ended questions might provide. As a result, they may not capture the full range of participant experiences or opinions, potentially oversimplifying complex viewpoints.

1.2.3.2 Unstructured Surveys

Open-ended questions are designed to allow participants to respond in their own words, fostering a more in-depth exploration of their thoughts and experiences. Examples of such questions include, "What are your thoughts on...?" or "Describe your experience with...?" One of the key advantages of using open-ended questions is that they provide deeper insights and richer data for qualitative analysis, enabling researchers to uncover unexpected themes or perspectives that may not have been anticipated. However, these questions also come with disadvantages; they can be time-consuming to analyze due to the variety of responses, making it challenging to draw consistent conclusions. Additionally, open-ended questions are susceptible to interviewer bias, especially when probing for further information, which may influence the direction of participants' responses.

1.2.4 Combining Survey Methods

In many cases, researchers opt for a mixed-methods approach by combining different survey methods or formats within a single survey. A survey might start with structured questions to gather basic demographic information, providing a quantitative foundation for the research. Following this, open-ended questions can be included to explore specific aspects of the topic in greater depth, allowing participants to express their thoughts and experiences more freely. By carefully selecting the survey type, format, and delivery method, researchers can enhance the accuracy and relevance of the data collected, ensuring it effectively addresses their research questions. This combination of quantitative and qualitative data allows for a more comprehensive understanding of the topic, enabling researchers to capture both statistical trends and personal insights.

Advantages of Surveys:

- Efficient way to collect data from a large sample.
- Relatively low cost compared to experiments.
- Can measure opinions, attitudes, and experiences.

Disadvantages of Surveys:

- Susceptible to bias - poor question wording, leading questions, or social desirability bias.
- Low response rates can affect the generalizability of results.
- Difficulty in verifying the truthfulness of responses.
- Cannot establish cause-and-effect relationships.

1.2.5 Experiments

Experiments are a powerful research method that allows researchers to test cause-and-effect relationships by manipulating one variable, known as the independent variable, and observing its impact on another variable, referred to as the dependent variable. Key components of experimental design include the control group and the experimental group. The control group does not receive the experimental treatment, serving as a baseline for comparison, while the experimental group receives the manipulation to assess its effect on the dependent variable. Randomization is also an essential aspect, as participants are ideally randomly assigned to either group to minimize bias and ensure that the groups are comparable.

The advantages of experiments include their status as the strongest method for establishing cause-and-effect relationships, as well as the ability to control extraneous variables that might otherwise confound the results. However, experiments also come with disadvantages. They can be expensive and time-consuming to set up and conduct, and ethical considerations may limit certain types of experiments, particularly those involving human or animal subjects. Additionally, the results obtained from experiments may not always generalize to real-world settings, raising questions about their external validity.

Let’ say we want to answer the following question.

Here's an example of an experiment for data collection designed to investigate the effect of fertilizer on plant growth:

Question: Does a specific fertilizer brand (Brand X) increase plant growth compared to a control group with no fertilizer?

In order to consider the use of fertilizer X can enhance plant growth, I would conduct a lab by setting two sets of pots and plant seeds of a plant in each group of pot. In the first set of pot, I would use the fertilizer X at an equal interval and in the second set of pots I would not use the fertilizer. I would keep everything the same but for use of fertilizer. I would then measure the growth of the plant at equal interval of time until the end of the period. That is, I would collect the data of the experimental group of a pot which I use the fertilizer and the control group of a pot for which I do not use the fertilizer. The two measures of the two sets of data are then used for the test of the hypotheses that the fertilizer X can increase the size of the plant. We will learn how to answer the above question later part of this book.

1.2.6 Observational Studies

Observational studies, unlike experiments, take a more passive approach to data collection. Imagine them as detailed observations of the world around us, rather than actively manipulating things to see what happens. Here's a breakdown of key points to elaborate on the concept:

- **No Manipulation:** Researchers don't control or introduce any changes to the variables they're studying. They simply observe and record what's already happening in a natural setting.
- **Focus on Associations:** Observational studies aim to identify relationships or associations between existing variables. A researcher might observe a link between smoking (independent variable) and lung cancer (dependent variable) in a large population, but they wouldn't force people to smoke to see if they get cancer.

1.2.7 Types of Observational Studies

- **Cross-Sectional Studies:** These studies capture data about a population at a single point in time. They're useful for understanding the prevalence of certain characteristics or associations between variables at that specific time. Imagine taking a snapshot of a population to see what's happening.
- **Longitudinal Studies:** These studies follow participants over time, repeatedly collecting data at multiple points. This allows researchers to observe changes and trends in variables, making them more powerful for suggesting cause-and-effect relationships compared to cross-sectional studies. Think of it as a movie capturing how things unfold over time.

1.2.7.1 Examples of Observational Studies

1. Studying the association between diet and heart disease by observing the eating habits and health outcomes of a large group of people over several years (longitudinal study).
2. Surveying students at a university to understand the prevalence of social media use and its impact on academic performance (cross-sectional study).

1.2.7.2 Advantages of Observational Studies

1. **Feasibility:** Studying real-world phenomena that cannot be easily manipulated in an experiment (e.g., the effect of social media on teenagers).
2. **Cost-Effective:** Often less expensive and time-consuming to conduct compared to experiments.
3. **Naturalistic:** Can provide insights into how variables interact in a natural setting.

1.2.7.3 Disadvantages of Observational Studies

1. **Confounding Variables:** Unobserved or uncontrolled factors that influence both the independent and dependent variables, making it difficult to establish a true cause-and-effect relationship.
2. **Direction of Causality:** Observational studies can only show correlations, not necessarily causation. Just because two things appear to be related doesn't necessarily mean one causes the other.
3. **Selection Bias:** The sample population may not be representative of the entire population of interest, leading to biased results.

By understanding the strengths and limitations of observational studies, researchers can leverage them to gain valuable insights into social phenomena, health trends, and other areas where manipulation of variables is not possible or ethical.

1.2.8 Sampling Methods

Sampling is a technique of choosing the sub-group (sample) from the larger group so that it represents entire population, for data collection purpose Sampling methods are statistical techniques used to select portion or subset of this "whole" (Population) in order to gain insight into its properties through taking measurements and making evaluations. Let us say you require the preferences of all the students in your university (population) about a new menu for one

cafeteria. It is not feasible to take a survey of every person. That means you take a sample of students that mirrors the overall student population, in your case through sampling method.

There are different sampling methods, each with its own advantages and disadvantages. The best method for your study depends on factors like:

- **The population:** Is it homogenous (similar throughout) or does it have subgroups (e.g., different age groups)?
- **Available resources:** Time, budget, and access to a list of the population.
- **Desired precision:** How important is it to have a perfectly representative sample?

In research, it's often impractical or impossible to study an entire population. This is where sampling comes in – a powerful technique for selecting a representative subset of the population to gather data. Here, we'll explore four common sampling methods: simple random, stratified, cluster, and systematic sampling.

1.2.9 Simple Random Sampling

The equivalent of drawing a name out of the hat for sampling is simple random sampling. If we can do that, in our minds eye we have a bowl full of tickets each representing one person from the population you are interested to study. In simple random sampling, each ticket is drawn with an equally likely chance - like a lottery. This is to make sure that each member of the population has an equal chance with one another to be selected into your sample. It is simple to implement—all you need is a way of classifying everyone into unique buckets and then selecting the correct number of participants with random numbers so skip ahead.

Pros of simple random sampling: This is the simplest type and widely used as well. One way this does is by reducing what biases can you get in, when everyone within a population has an equal

opportunity for being chosen. It prevents over-representation of bias in some subpopulations and makes the result more population-neutral. It fosters diversity. Random selection of who got to participate in the study makes it more likely that your sample will represent all the characteristics present in the overall population and thus findings are considered generalizable. In other words, what you are learning from the sample can be definitely generalized to most of that group. It is a simple technique to apply for many research projects, as assigning unique identification numbers and using the random number generator goes quite quickly. But it does have some limitations. Although unbiased when performed correctly, simple random sampling may not always represent the full diversity of a population if it has clearly delineated subgroups. This form of sampling might not always yield a sample that is truly representative, especially if the population under consideration isn't homogeneous.

1.2.10 Stratified Sampling

A stratified sample type includes different subgroups in a data set to make it representative (like slicing your cake into layers so each flavour gets a nod). Analyse the population and split it into groups (strata) based on characteristics that might affect a particular behaviour or consumption, such as age, gender, and income. There is a more complex approach to it. Next, each stratum is randomly sampled to put a skin in the game for all subgroups. This is especially helpful when the population consists of clear subgroups you want to ensure are present in your final sample.

Let's say a researcher is interested in the movie preferences of teenagers (those between 13 and

21) within a metropolitan area. The city has a wide-ranging, cosmopolitan population, and teens of different generations will likely each have their own set of must-see movies. To capture these possible variations, the researcher may use stratified sampling to secure a representative sample. To do this, they would first stratify the teenage population based on age groups (e. g., 13-15, 16-18 and so forth.). They would then determine what percentage of the teenager population was each age group. Lastly, they would perform probability simple random sampling for each age group by selecting a size-based number of participants from the population. This approach guarantees that both 13 year old children, all the way up to 17 year old and everyone in between has a chance of being represented within the sample which provides you with an insight into preferences for different ages across middle-teenagers.

Stratified sampling is alluring because you can ensure yourself that essential subgroups within your studies are represented, as opposed to simple random sampling. This is like making sure there is a piece of the cake for each subgroup in the population, similar to migrations, where our migration hit all possible subgroups. Stratified sampling involves splitting the population into similar groups (strata), related to specific characteristics, like age, gender or income level. This is especially useful when you already know some subgroups are going to be important for your study, and you need their voices in the conversation. How are stratified samples different from simple random sample? This results in stronger and generalizable conclusions, as the resulting sample is then more representative of these stratification elements within a population.

On the other hand, although stratified sampling is efficient in terms of subpopulation representation it has its own disadvantages. The big barrier is the first step: defining meaningful strata. This relies on pre-existing information of the population and a list of traits that could potentially affect the dependent variable. Selecting the strata that you use wrongly or ignoring significant subgroups resulting in biased goodness of fit. Furthermore, stratified sampling can take more work to implement than simple random sampling. Splitting the population, measuring each subgroup proportional with a random sample from these strata takes more upfront work and consideration to implement. Finally, since this cannot be the

most viable sampling approach if there are some strata which is altogether small or difficult to get a hand on - then it may influence how overall representative of your sample will turn out. Nevertheless, stratified sampling continues to be useful when it is absolutely critical that subgroups are represented in sufficient numbers for a study's success.

1.2.11 Cluster Sampling

Cluster sampling is a method in which the population is first divided into several cluster. Then at random a set of clusters is selected. Finally all units in the selected clusters are: included in the sample. This similar to select a sample from the entire state. Then divide the state into different districts and then randomly select some of the districts. Finally all members in the selected districts are included in the sample.

It is a cheap way to take a large sample without a field survey as is necessary in case of random sampling. Cluster sampling is a clever way of sampling for situations where you can't possibly track down each individual in a large, dispersed population. Here's an example to explain this method. Let's say you want to study the eating habits of people residing in far-flung villages. Attempting to talk to each person in this case would be a logistical nightmare. So, what do you do? Use cluster sampling. The villages here will act as a naturally occurring group, known as a cluster, which will ideally be a miniature version of the entire population. Moreover, this method is not only efficient but also pocket-friendly. However, as it often happens with such smart solutions, there's a caveat. The clusters should be representative of the true population.

In other words, if all the villages eat the same kind of food, there's no problem. However, if the eating habit is significantly different from one village to another, the sample might not yield accurate results regarding the larger population. It's a little bit like forming an opinion of an entire forest based on specific trees.

Cluster sampling is advantageous in terms of strategy when the population of interest is scattered across a geographical area or when it is difficult to reach individual population members. To illustrate, consider the case of data examination to understand the attitudes and preferences of farmers in a rural area, for example. First and foremost, it is complicated and protracted to trace and address each farmer from near and far. Secondly, it will be expensive because it may involve many resources. Thus, cluster sampling `rescues' the researcher. With this aid, the entire population is divided into groups or clusters, which, in an ideal case, repeat the structure of the population. Whichever clusters take place, the researcher should randomly make a choice. Finally, one is only to address people in the selected groups. The merit of this type is evident: it is cheaper and faster. What is more, it becomes the only possible solution when the funds are limited, or there is no possibility of tracing individual representatives of the population. The demerit of the selected membership type is directly related to the selection of clusters. Being connected is the defining disadvantage if, in the case with farmers, all representatives were vegetable farmers. It is beneficial in terms of comparison, but in this case, the comparative estimating of all products within the whole territ9ry will not take place.

Apart from the aforementioned demerits, this method may also lead to increased sampling error. The matter is that all clusters may be homogeneous. All villages grow rice. The third and last demerit is that the strategy becomes less informative. The only one way to obtain additional information is to select the other aggregate or district to be processed. Nevertheless, this method is still considered fairly safe and reasonable under the circumstances mentioned above.

1.2.12 Systematic Sampling

One should imagine lining up all the members of the population and then picking every nth member for the sample. This is how systematic sampling works. First, the population is put in a list, and then a random starting point is figured out. Finally, every nth member after the starting point is included in the sample. Thus, the structure of systematic sampling is more or less intuitive with the work done at each step being relatively straightforward. Therefore, imagine that all your class students are lined up alphabetically, and you want to survey every nth student about their study habits. You will have to use systematic sampling. First of all, all students must be put into a list. Most likely, the students were arranged based on their names that could be later alphabetically ordered. Then, a random starting point on that list should be determined. This could become any number between 1 and the sampling interval, which is the total number of students divided by the desired sample size. Finally, one will have to count every nth student on the list, starting from the chosen point and include them in the sample. The biggest advantage of this method is that it is simple and efficient. However, it depends on the original list to be truly random or at least not hiding any other patterns. For example, a student can be accidentally selected in the sample if they are born in a particular month and are grouped together in the list. Systematic sampling could be a preferred technique if one has an ordered and complete list of a population, but one should be cautious about potential ordering-caused biases.

Systematic sampling is the statistical technique that strikes the balance between simplicity and efficiency. If everyone in the population can be lined in a row, one can pick every nth person for the survey, no more no less. The main idea of systematic sampling is using order. First of all, the population should be listed and there are order principles. For example, movie-goers

can be listed by arrival, by names, by ticket numbers. After ordering, one should select the first element randomly and then count n-1 elements to capture the one for the survey. The great advantages of systematic sampling are simplicity and efficiency. Once the list is ready, there is no need to look for members one by one. However, the disadvantage is also system, as the result totally depends on the initial order. If the order is orderly non-random, i.e. movie-goers ordered by the groups, and the youngest members are before the oldest in the list, the systematic sample will include more young people and will represent better the younger generation. Another disadvantage is potential systematic error: certain groups will be counted together. The list can include similar people with the group of children occupying 1 place out of 7. The sample will be less representative, because 6 out of 7 will be other children and the 7th is adults. On the other hand, systematic extraction is a perfect solution if you have an ordered list for several reasons: the algorithm will be better ordered and will take less time and efforts.

Recap

- **Data Importance**: Reliable data ensures conclusive research results.

- ***Main Data Collection Methods***

1. **Surveys**: Gather data using pre-set questions.
    - Telephone: Cost-effective but limited by mobile phone use.
    - Face-to-Face: Detailed, labor-intensive.
    - Online: Fast, broad reach, risk of bias.
    - Mail: Traditional, declining use.
    - SMS: Quick and targeted.
    - Kiosk: Convenient in public spaces.
    - Formats:
- Structured: Easy to analyze, limited depth.
- Unstructured: Rich data, hard to analyze.
    - Mixed methods combine quantitative and qualitative data.
    - **Advantages**: Cost-efficient, measures opinions.
    - **Disadvantages**: Biases, low response rates, no causation.
2. **Experiments**: Test cause-and-effect relationships.
    - **Components**: Control group, experimental group, randomization.
    - **Advantages**: Best for causation, controls variables.
    - **Disadvantages**: Costly, ethical limits, limited generalizability.
    - **Example**: Testing fertilizer effects on plant growth.
3. **Observational Studies**: Observe variables without manipulation.
    - Types:
- Cross-Sectional: One-time data (e.g., social media use).
- Longitudinal: Over time (e.g., diet and heart disease).
    - **Advantages**: Feasible, cost-effective, naturalistic.
    - **Disadvantages**: Confounding factors, no causation, selection bias.

- ***Sampling Methods***

- **Simple Random**: Equal chance for all, less diverse.
- **Stratified**: Ensures subgroup representation, needs pre-data.
- **Cluster**: Cost-efficient, higher error risks.
- **Systematic**: Simple, risk of ordering bias.

- ***Choosing Methods***
  - Match to population, budget, and study goals.

Objective Type Questions

1. What is the main purpose of conducting experiments in research?
2. Define the independent and dependent variables in an experiment.
3. What role does a control group play in an experimental design?
4. Explain the importance of randomization in experiments.
5. List an advantage of using experiments as a research method.
6. How do observational studies differ from experiments in data collection?
7. Distinguish between cross-sectional studies and longitudinal studies.
8. List an advantage of observational studies.
9. Describe the concept of sampling in research.
10. How does simple random sampling ensure unbiased results?
11. Explain the process of stratified sampling.
12. What is systematic sampling?

Answer

1. Experiments allow researchers to test cause-and-effect relationships.
2. The independent variable is manipulated in an experiment and dependent variable is observed in an experiment.
3. A control group does not receive the experimental treatment and experimental group receives the manipulation.
4. Randomization minimizes bias in experimental design.
5. Experiments are considered the strongest method for establishing cause-and-effect relationships.
6. Observational studies do not involve manipulating variables.
7. Cross-sectional studies capture data at a single point in time and longitudinal studies follow participants over time to observe changes.
8. Observational studies can identify relationships between existing variables.

9. Sampling is a technique for selecting a representative sub-group from a larger population.
10. Simple random sampling gives every member of the population an equal chance of selection.
11. Stratified sampling ensures representation of essential subgroups within a study.
12. Systematic sampling involves selecting every nth member from an ordered list.

Suggested reading

1. Cochran, William G. *Sampling Techniques*. 3rd ed., Wiley, 1977.
2. Moore, David S., et al. *Introduction to the Practice of Statistics*. 9th ed., W.H. Freeman, 2017.
3. Newbold, Paul, et al. *Statistics for Business and Economics*. 8th ed., Pearson, 2013.
4. Sekaran, Uma, and Roger Bougie. *Research Methods for Business: A Skill-Building Approach*. 7th ed., Wiley, 2016.
5. "Sampling Methods." *Statistics How To*, www.statisticshowto.com/probability-and-statistics/sampling-methods/. Accessed 15 Nov. 2024.
6. "Study Design: Sampling Methods." *Khan Academy*, www.khanacademy.org/math/statistics-probability/designing-studies/sampling-methods. Accessed 15 Nov. 2024.
7. "Survey Sampling and Estimation." *National Center for Biotechnology Information*, www.ncbi.nlm.nih.gov/books/NBK305072/. Accessed 15 Nov. 2024.

UNIT -3

Types of Data

Learning outcome

After the successful completion of the course, the learner will be able to:

- Define qualitative and quantitative data.
- Identify examples of quantitative data types.
- List examples of qualitative data collection methods.
- Recognize methods for analysing qualitative data.
- Describe the four levels of measurement: nominal, ordinal, interval, and ratio.

Pre-requisites

Understanding qualitative and quantitative data is essential for effective data analysis. Qualitative data, which is descriptive and subjective, is analyzed using methods like thematic and narrative analysis to identify patterns and meanings. Quantitative data, which is numerical and objective, can be continuous or discrete and is analyzed using statistical tests. The levels of measurement—nominal, ordinal, interval, and ratio—determine the appropriate statistical techniques, with each level offering different analytical possibilities. Scaling techniques such as Likert and semantic differential scales are used to quantify attitudes and perceptions. Mastery of these concepts allows researchers to design robust studies, perform comprehensive analyses, and draw meaningful conclusions from diverse data sets.

Keywords

Qualitative Data, Quantitative Data, Continuous Data, Discrete Data, Measurement- Nominal, Ordinal, Interval, Ratio Measurement

Discussion

1.3.1 Introduction to types of data

Understanding the types of data is fundamental in data analysis. Data can be broadly classified into two categories: qualitative and quantitative. Each type of data has distinct characteristics and requires different methods of analysis.

1.3.2 Quantitative data

Quantitative data can be defined as the information that can be measured and expressed with numbers. It is used to quantify variables, similarly to the nominal data, but provides more precise and objective measurements that can be used for statistical analysis and various mathematical computations. There are several methods of collecting this type of data. Some of the most common among them are surveys with closed-ended questions, experiments, as well as secondary data analysis of records and databases other people or organizations leave. It can be both discrete, counting items, such as the people in a room or the cars on the street, and continuous, measuring the data on a continuous scale, such as human height, temperature, distance, or blood pressure. Some examples of quantitative data are the results of surveys that analyze respondents from zero to ten or from one to seven, various financial and budget data on sales, revenue, cost, or market share, as well as test scores, demographic data, such as age, revenue, and population, and many others. The data analysis process, in this case, may contain such methods as descriptive statistics, to summarize the data, inferential statistics, to make predictions and/or inferences with regard to the bigger population, and various techniques, such as correlation and regression, used to research relationships between variables. This type of data is used for empirical research, which allows for testing hypotheses, identifying trends, and making reasonable decisions based on data.

Examples:

- **Test Scores:** Quantitative measures of performance or ability.
- **Financial Data:** Numerical records of sales, revenue, costs, and other financial metrics.
- **Demographic Data:** Statistics such as age, income, number of siblings, and other measurable attributes.
- **Experimental Data:** Measurements such as reaction times, temperature readings, and other controlled variable observations.

1.3.2.1 Advantages of Quantitative Data

- **Objectivity and Precision:** Quantitative data is numerical, which provides precise and objective measurements. This allows for clear, unambiguous interpretation of results.
- **Statistical Analysis:** Quantitative data can be analyzed using a wide range of statistical techniques, from basic descriptive statistics to complex inferential analyses. This facilitates hypothesis testing, trend analysis, and predictive modeling.
- **Generalizability:** When collected from a representative sample, quantitative data allows researchers to generalize findings to a larger population. This is particularly useful for making broad claims and predictions.

- **Reproducibility:** Quantitative research methods are typically standardized, which means that studies can be replicated by other researchers to verify results. This enhances the reliability and validity of the findings.
- **Comparability:** Numerical data allows for easy comparison across different groups, time periods, or variables. This makes it possible to identify patterns, correlations, and differences with a high degree of accuracy.
- **Efficiency:** Quantitative data collection methods, such as surveys and experiments, can be conducted relatively quickly and can handle large amounts of data efficiently. This is advantageous for studies that require analysis of extensive datasets.
- **Visual Representation:** Quantitative data can be easily represented in various graphical formats (e.g., charts, graphs, tables), making it accessible and understandable for a wide audience, including those who may not be familiar with the underlying data.
- **Decision-Making:** Quantitative data provides a strong empirical foundation for making informed decisions in various fields, including business, healthcare, and public policy. The numerical nature of the data supports evidence-based decision-making and strategic planning.
- **Automation and Technology:** With advancements in technology, quantitative data can be collected, stored, and analyzed using automated tools and software. This increases efficiency and reduces the potential for human error in data handling and analysis.

1.3.3 Qualitative data

In simple words, qualitative data is a type of data that is not likely numerical. It is the type of data which describes a thing with quantity using qualities, frequencies, or attributes, among others. It provides more depth to research, especially its context, and it is usually a descriptive type. In sociology, qualitative data is data that has to do with being studied or existing in which the rate of description cannot be determined. It is not based on numbers and is gathered through a variety of methods, including interviews, observations, and focus groups, which can be either formal or informal. Qualitative data is used to establish the social phenomena and get a range of information from the social aspects and opinions of others. Detailed interview scripts, examination records, field notes, open-ended survey responses, and other resources come under more qualitative data. What makes qualitative data so important and rich is that it can reveal certain patterns and themes, which would not be revealed by the quantitative methods, and it is much more in-depth and thorough. Thematic, content, and narrative analysis are the most common analytical methods of research. This sort of data should be taken into account in every field that does research.

An example of qualitative data is descriptions of customer experiences with a new product. It can be obtained through open questions in a survey, an interview, or a focus group. The customers express their attitudes, emotions, and opinions about the product they have tried. For example, it can be “I love the sleek design and user-friendly interface,” or “The product is too expensive for the features it offers.” It is the kind of information that can give businesses the perception and preference or criticism of their product or its characteristics. In such a way, the businesses can learn what exactly is appealing to their customers or which things they need to improve.

Examples of Qualitative Data:

1. Interview Transcripts:
    - Detailed records of conversations with individuals where they share their experiences, opinions, and perspectives.
    - Example: An interview transcript from a study exploring the experiences of cancer patients, where participants describe their feelings, challenges, and coping mechanisms.
2. Open-Ended Survey Responses:
    - Written responses to survey questions that allow participants to express their thoughts and feelings freely.
    - Example: Responses to a customer satisfaction survey asking, "What do you like most about our product?" and "How can we improve our service?"
3. Field Notes:
    - Observational notes recorded by researchers during or after an event or activity.
    - Example: A researcher's field notes from observing classroom interactions in a kindergarten, documenting student behaviors, teacher-student interactions, and classroom dynamics.
4. Photographs and Videos:
    - Visual documentation that captures events, settings, or subjects.
    - Example: Photographs from a community event showcasing different activities, participant interactions, and the overall atmosphere. Videos of wildlife behavior recorded during a field study.
5. Text Documents:
    - Existing texts such as books, articles, reports, and other written materials.
    - Example: Analysis of newspaper articles covering a particular political event to understand media portrayal and public sentiment. Examination of historical letters to study communication styles and cultural norms of a specific period.
6. Focus Group Discussions:
    - Transcripts from group discussions guided by a facilitator where participants discuss their views on a specific topic.
    - Example: A focus group discussion with parents about their experiences and concerns with remote learning during the COVID-19 pandemic.
7. Case Studies:
    - In-depth analysis of a single case or a small number of cases.
    - Example: A case study of a successful startup, including interviews with the founders, employees, and customers, as well as analysis of company documents and market strategies.
8. Diaries and Journals:
    - Personal records kept by individuals, documenting their daily activities, thoughts, and feelings.
    - Example: Diaries of teenagers documenting their daily life and emotional experiences for a study on adolescent mental health.

9. Social Media Posts:
    - Content posted on social media platforms, including text, images, and videos.
    - Example: Analyzing Twitter posts related to a natural disaster to understand public reactions, concerns, and the spread of information.
10. Ethnographic Records:
    - Detailed descriptions and analyses from ethnographic studies where researchers immerse themselves in a community or culture.
    - Example: Ethnographic records of a remote indigenous community, including observations of daily life, rituals, and social interactions.

These examples of qualitative data provide rich, detailed insights that help researchers understand the complexities of human behavior, social interactions, and cultural phenomena. By analyzing this data, researchers can uncover patterns, themes, and meanings that quantitative data alone cannot reveal.

1.3.4 Analysis of Qualitative Data

Analysis of qualitative data implies understanding non-quantifiable pieces of information, such as meanings and patterns. Some of the methods of doing that are specific to types of research questions and types of data. There is a common thematic analysis method where a researcher has to find the text, read it and get familiar with the data, code the text in connection to the theme, accumulate the data and re-read them to determine the theme, review the themes, and finally write the report. The thematic analysis method is relatively flexible and is used in various fields. The other method is content analysis, which is based on detailed analysis, coding, and categorizing of texts to identify implications, meaning, and patterns. This is also a common method and is used to quantify the presence of subjects at the intersection of media content, documents, and symposiums.

Another commonly used method is discourse analysis that involves written and spoken words with their social context. This method helps to examine how the language constructs the social realities and customs. One of the types of discourse analysis, narrative analysis, implies examining the stories of people to explain how they make sense of their experiences. Stone explores this method in relation to the analysis of Iraq war letters. The grounded theory is directly connected with data analysis for theory building. This method is based on developing the theory with the help of data, in inductive form. Phenomenological analysis involves studying the perceptions and experiences of the people with the aim to construct a description of a phenomenon. Framework analysis is much more systematic in analyzing and managing the data in the already available theoretical frameworks. Case study analysis implies looking at one or several cases within the real-life setting. Finally, constant comparative analysis is generally used in grounded theory and involves a continuous comparison of data chunks. Another method to be discussed, Interpretative Phenomenological Analysis emphasizes the significance of making sense that people put into their personal worlds. All these methods can be used for qualitative data analysis to help work with the data in distinct ways and make unique contributions to understanding.

1.3.5 Converting Qualitative Data to Quantitative Data

Converting qualitative data to quantitative data involves transforming non-numerical, descriptive information into numerical values that can be analyzed statistically. This process is essential for integrating qualitative insights into quantitative frameworks, enabling a more comprehensive analysis. Here are some common methods to achieve this conversion:

1.3.5.1. Coding and Categorizing

1. Thematic Coding

Qualitative data analysis involves identifying key themes or concepts from sources such as interview transcripts or open-ended survey responses. The process begins by carefully reading through the data and highlighting significant phrases. These phrases are then assigned short labels, known as codes. Similar codes are grouped to form broader themes, providing structured insights. To add a quantitative aspect, the frequency of each code or theme is counted, converting qualitative information into numerical data. For example, in a customer satisfaction study, responses like “excellent service” or “quick delivery” can be coded and analyzed based on their occurrence.

2. Content Analysis

Content analysis involves systematically categorizing textual information into predefined categories to identify patterns and trends. The process begins by developing a coding scheme with specific categories relevant to the analysis. Each piece of data is then assigned to the appropriate category based on its content. Once categorized, the frequency of occurrences is counted, allowing for statistical analysis. This approach helps transform qualitative data into measurable insights. For example, in an analysis of news articles, topics such as "economic issues," "healthcare," or "politics" can be coded and quantified to assess their prominence.

3. Rating and Scaling

- Likert Scales: The process of transforming qualitative responses into numerical ratings involves using a structured scale, such as a Likert scale, to measure opinions or perceptions. First, key themes from qualitative data are identified, and relevant statements are formulated. Respondents then rate their agreement or satisfaction on a predefined scale, typically ranging from 1 (strongly disagree) to 5 (strongly agree). The collected ratings are analyzed using statistical measures like mean, median, or mode to identify trends. For example, customer feedback on service quality can be assessed by averaging ratings for statements like "The customer service was helpful."

- Semantic Differential Scales: The process of using bipolar adjectives involves converting qualitative perceptions into numerical data by presenting respondents with a scale featuring opposite adjectives. A scale is designed with pairs like "satisfied" vs. "dissatisfied" or "easy to use" vs. "difficult to use," allowing respondents to indicate their position between the two extremes. The collected ratings are analyzed to determine central tendency (mean, median, mode) and dispersion (range, standard deviation). For example, a product's usability can be assessed by asking users to rate it on a scale from 1 ("easy to use") to 7 ("difficult to use").

4. Quantitative Content Analysis

Word Frequency Analysis: Word frequency analysis involves counting the occurrence of specific words or phrases in qualitative data to identify patterns. Text analysis software is used to scan the data, extract key words, and count their frequency. The numerical data generated can then be analyzed statistically to determine trends or common themes. For example, in analyzing social media comments, words like "happy," "disappointed," or "recommend" can be counted to understand customer sentiment and overall perception.

5. Sentiment Analysis: Sentiment analysis involves categorizing qualitative data into positive, negative, or neutral sentiments. This process typically uses sentiment analysis tools or algorithms to classify each response based on its emotional tone. The results are then quantified by calculating the proportion of each sentiment category. For example, in customer reviews, the percentage of positive, negative, and neutral feedback can be determined to assess overall customer satisfaction and brand perception.

Qualitative to quantitative data conversion is the practice that refers to the use of a system of systematic procedures which implies the transformation of descriptive data into information described in terms of figures. As a methodological tool that helps combine rich and comprehensive information associated with context and statistics, the approach involves techniques, such as thematic coding, categorizing, content analysis, and scaling. As the result of the investigation, the obtained information reveals valuable information for the process of making decisions, identifying the correlations, and conducting further research.

1.3.6 Levels of Measurement

Understanding the levels of measurements is essential when it comes to data analysis. Accordingly, there are four types of measurements that can be relevant, that include nominal, ordinal, interval, and ratio measurements. Each type is concerned with the determination of the nature of data, as well as the operations that could be applied to the data. The nominal type is mostly used for the classification of data into relatively equal categories. Usually, there is no clear order of the categories. Examples include gender, marital status, and nationality. The primary operations that could be performed with the nominal data are counting frequencies and using modes. As for statistical data analyses, the appropriate test of significance is a chi-square test, which could be applied to determine the relationship or association between the two or more categorical variables.

In order to increase the number of statistical techniques that can be used, nominal measurement is the first level of measurement, referring to categories that cannot be put into any meaningful order, thus permitting few statistical analyses. Examples include sex, gender, religion, and occupation. In contrast, ordinal measurement is the next level, referring to categories that can be ordered or ranked, but the intervals between such ranks are not equal nor known. Examples are education levels, socioeconomic status, and satisfaction ratings. While the statistical measures of central tendency that will be appropriate for ordinal data are the median and mode, advanced statistical tests beyond descriptive statistics involve non-parametric tests such as the Mann-Whitney U test. After ordinal measurement is interval measurement, which includes meaningful intervals between any ordinally-ordered category, for comparison of differences. Examples include all temperature scales and IQ scores. However, like the first two levels, it cannot have a true zero point, making ratios and their interpretation meaningless. Finally, ratio measurement is the most informative level of measurement, which always has a true zero point that enables all arithmetic operations including ratios. Examples include height, weight, and income. For the latter two levels, comprehensive descriptive statistics and inferential statistics

can be applied, such as mean, standard deviation, t-test, and ANOVA. Therefore, recognizing such levels can help researchers develop appropriate methods for data collection, analysis, and interpretation to create more valid and reliable research.

1.3.6.1 Nominal Level

The nominal level of measurement is the most basic form of data categorization, where variables are classified into distinct, mutually exclusive categories without any intrinsic order or ranking. Each category is unique and serves merely as a label or identifier for different attributes or groups. Variables such as gender (male, female, non-binary), marital status (single, married, divorced, widowed), and nationality (American, Canadian, British, etc.) fall under the nominal level. Since these categories do not have a logical sequence, the primary operations involve counting frequencies and determining the mode, which is the most common category. Nominal data is often analyzed using chi-square tests to examine relationships between categorical variables. This level of measurement is crucial for classifying and organizing data in fields such as demographics, marketing, and social sciences.

Characteristics:

- Categories are mutually exclusive and collectively exhaustive.
- No logical order or ranking among the categories.

Examples:

- Gender (male, female, non-binary)
- Marital status (single, married, divorced, widowed)
- Blood type (A, B, AB, O)
- Nationality (American, Canadian, British, etc.)

1.3.6.2. Ordinal Level

The ordinal level of measurement is concerned with partitioning of data into separate categories, which can be ranked, ordered, and demonstrating different place ranges on the scale, but the intervals between these categories are unequal or unknown. This method allows for an indication of relative positions, such that we can determine whether one item is more or less than another or better or worse, but the scale does not permit mathematical operations such as addition or subtraction. Some examples of measures to be counted using ordinal scale are levels of education, socioeconomic status, and continuous levels of satisfaction. Ordinal scale provides an opportunity to obtain data where the order of the placement is meaningful, but the precise difference between values is not known. While measuring ordinal data, it is recommended to apply non-parametric measures such as median, rank-based test, etc., rather than measures which specify uniform intervals between their units, or parametric measures.

Example of Ordinal Level of Measurement

Consider a customer satisfaction survey for a new software product. Respondents are asked to rate their overall satisfaction on a five-point scale:

1. Very Dissatisfied
2. Dissatisfied
3. Neutral
4. Satisfied
5. Very Satisfied

In this scenario, each response falls into a distinct category that can be ordered from least satisfied to most satisfied. This ranking reflects an ordinal level of measurement because it allows us to determine the relative position of each response. For example, we know that "Very Satisfied" is better than "Satisfied" and "Satisfied" is better than "Neutral." However, the exact difference in satisfaction between each category is not quantifiable.

Frequency Distribution: Suppose the survey results are as follows:

- Very Dissatisfied: 5 responses
- Dissatisfied: 10 responses
- Neutral: 15 responses
- Satisfied: 25 responses
- Very Satisfied: 45 responses

The frequency distribution indicates the number of respondents in each category, showing the overall trend in customer satisfaction.

To find the median satisfaction level:

1. List all responses in order: 5 (Very Dissatisfied), 10 (Dissatisfied), 15 (Neutral), 25 (Satisfied), 45 (Very Satisfied).
2. Identify the middle value. With 100 total responses, the median is the 50th value, which falls in the "Satisfied" category.

To compare satisfaction levels across different demographic groups, non-parametric tests like the Mann-Whitney U test or the Kruskal-Wallis test can be used. These do not assume equal intervals between categories and are suitable for ordinal data. This example illustrates how ordinal data is used to rank and analyze categorical responses, emphasizing the order without assuming equal spacing between ranks.

1.3.6.3. Interval Level

The interval level of measurement is defined as a numerical scale in which the intervals of values are consistent and meaningful. One of the major defining characteristics of the interval scale is the lack of a true zero. Meaning that it is possible to add and subtract the values within the scale, though finding ratios is not possible. The examples of such systems in nature are Celsius and Fahrenheit temperature scales, in which a difference between every degree is equal but zero does not represent the lack of temperature. An arbitrary example of an interval level of measurement is the IQ score, in which a value of zero means nothing. The mean, median, and standard deviation can be calculated at the interval level of measurement, which makes it useful for all kinds of scientific, social, and psychological research. One can find an average temperature for a certain month and then compare the difference in temperature between two months. However, one cannot say that 40 degrees is twice as hot as 20 degrees. Unlike with the ratio level of measurement, such comparison loses justification at the ratio level of measurement, which is why one can say they offer an intuitive limitation of this level of measurement. Knowing that the level is interval helps to draw understanding of the interval level of measurement is useful and should be used when research implications require understanding of statistics, while the variables involved should be strictly based on the interval level of measurement.

Example of Interval Level of Measurement

Consider measuring temperature using the Celsius scale. This scale is a perfect example of the interval level of measurement, where the difference between each degree is consistent and

meaningful. The difference between 20°C and 30°C is the same as the difference between 30°C and 40°C; both intervals represent a change of 10 degrees.

Mean Calculation: Suppose we have the following temperature readings over a week: 22°C, 25°C, 27°C, 23°C, 26°C, 28°C, and 24°C. To find the average temperature for the week, we sum these values and divide by the number of readings:

$$\text{Mean} = \frac{22 + 25 + 27 + 23 + 26 + 28 + 24}{7} = \frac{175}{7} = 25°C$$

Standard Deviation Calculation: To understand the variability in temperature, we calculate the standard deviation. The steps include finding the mean, subtracting the mean from each reading to get the deviations, squaring these deviations, finding the average of the squared deviations, and taking the square root of this average.

1. Mean: 25°C
2. Deviations: -3, 0, 2, -2, 1, 3, -1
1. Squared Deviations: 9, 0, 4, 4, 1, 9, 1
2. Average of Squared Deviations (Variance):
$$\frac{9 + 0 + 4 + 4 + 1 + 9 + 1}{7} = \frac{28}{7} = 4$$
3. Standard Deviation: $=\sqrt{4} = 2°C$

This calculation shows the average deviation from the mean temperature, giving us a sense of the variability in daily temperatures.

Limitations: While we can perform these operations, the Celsius scale does not allow for meaningful ratio comparisons. For example, saying that 40°C is twice as hot as 20°C is incorrect because the scale lacks a true zero point where there is an absence of temperature.

Practical Application: This interval scale is essential in meteorology for weather forecasting, where precise differences in temperature readings inform predictions and climate studies. Researchers and meteorologists use these consistent intervals to compare and analyze temperature changes over time and across different regions, enhancing our understanding of weather patterns and climate change.

This example highlights the strengths and limitations of interval data, demonstrating its utility in statistical analysis while cautioning against improper ratio interpretations.

1.3.6.4. Ratio level

The ratio level of measurement is the most informative and most precise of all the level of measurements since it has all the properties of the interval level plus an absolute or true zero. It refers to a value that means a complete absence of having the quantity measured. It can be expressed in terms of a full range of mathematical operations, including meaningful addition, subtraction, multiplication, and division, thus can be related to each other regarding the prospect of having less or more of the quantity being measured. Examples of ratio data are height, weight, age, income, and distance. Being twice the height of 120 cm, for example, is clearly meaningful. Thus, the standard deviation and the range can be calculated, and the ratios between measurements in the same variable can be meaningfully taken. This level of measurement is typical for the following fields: physical sciences, economics, and engineering. In conclusion, the ratio level of measurement is the most accurate, informative and best, that

allows a researcher to analyze data.

1. **Example of Ratio Level of Measurement**

Consider measuring the weights of a group of athletes in kilograms. This scenario exemplifies the ratio level of measurement, as it has all the properties of interval measurement with an absolute zero point, which indicates the absence of weight.

2. **Data Collection**

Suppose we have the following weights for a group of athletes: 60 kg, 70 kg, 80 kg, 90 kg, and 100 kg.

3. **Mean Calculation:** To find the average weight, we sum the weights and divide by the number of athletes:

$$\text{Mean} = \frac{60 + 70 + 80 + 90 + 100}{5} = \frac{400}{5} = 80 \text{ kg}$$

4. **Standard Deviation Calculation:** To understand the variability in weights, we calculate the standard deviation. The steps include finding the mean, subtracting the mean from each weight to get the deviations, squaring these deviations, finding the average of the squared deviations, and taking the square root of this average.

1. Mean: 80 kg
2. Deviations: -20, -10, 0, 10, 20
3. Squared Deviations: 400, 100, 0, 100, 400
4. Variance (Average of Squared Deviations):
$$\frac{400 + 100 + 0 + 100 + 400}{5} = \frac{1000}{5} = 200$$
5. Standard Deviation:

$$\sqrt{200} = 14.14$$

Ratio Comparisons: With ratio data, we can make meaningful ratio comparisons. For example:

- An athlete weighing 100 kg is twice as heavy as an athlete weighing 50 kg.
- The difference in weight between athletes (e.g., 100 kg - 60 kg = 40 kg) is meaningful and can be used in further calculations such as total team weight.

Practical Application: In sports science, understanding the weights of athletes can help in assessing health and performance metrics. For example, trainers can calculate the body mass index (BMI) of athletes to monitor their fitness levels, design diet plans, and tailor training regimens. The ratio level of measurement allows for precise and comprehensive analysis, supporting data-driven decision-making in sports and many other fields.

This example demonstrates the strengths of ratio data, emphasizing its ability to support a wide range of mathematical and statistical operations while providing meaningful and interpretable results.

Recap

- Quantitative data is information that can be measured and expressed numerically.
- Advantages of Quantitative Data:
  - ➢ Objectivity and Precision:
  - ➢ Statistical Analysis
  - ➢ Generalizability:
  - ➢ Reproducibility:
  - ➢ Comparability
  - ➢ Automation and Technology:
- Qualitative data is non-numerical and descriptive, providing depth to research, especially in its context.
- Examples of Qualitative Data:

➢ Open-Ended Survey Responses: Written responses expressing thoughts freely.
➢ Field Notes: Observational notes recorded during events.
➢ Photographs and Videos: Visual documentation of events.
➢ Text Documents: Analysis of books, articles, and reports.
➢ Focus Group Discussions: Group discussions on specific topics.

- Levels of Measurement
  - ➢ Nominal Level
  - ➢ Ordinal Level
  - ➢ Interval Level
  - ➢ Ratio Level

Objective Questions

1. What are the two broad categories of data?
2. What type of data is expressed with numbers?
3. What type of quantitative data measures on a continuous scale?
4. What type of data can include test scores and demographic data?
5. What level of measurement uses mutually exclusive categories without order?
6. What is an example of nominal data?
7. What level of measurement has ordered categories with unequal intervals?
8. What is an example of ordinal data?
9. Which level of measurement has consistent intervals but no true zero?

Answers

1. Qualitative
2. Quantitative
3. Continuous
4. Quantitative
5. Nominal
6. Gender
7. Ordinal
8. Education
9. Interval

Assignment Questions

1. Compare qualitative and quantitative data in terms of their characteristics and applications. Provide examples to illustrate your points.
2. Explain the process of thematic analysis in qualitative research. How does it differ from content analysis?
3. Explain the four levels of measurement (nominal, ordinal, interval, and ratio) and provide examples for each. Discuss the implications of each level for statistical analysis.
4. Describe the semantic differential scaling technique and its applications. How does it differ from Likert scaling?
5. Discuss the appropriate statistical tests for analysing nominal and ordinal data. Provide examples of scenarios where each test would be used.

Suggested reading

1. K Huffman and R Kunze, Linear Algebra, Pearson Education, 2nd Edition, 2005.
2. Thomas M. Apostol, Calculus, Wiley, 2nd Edition, 1991 ISBN 960-07-0067-2.
3. Michael Spivak. Calculus, publish or Perish, 2008, ISBN 978-0914098911.
4. Ross L. Finney, Maurice D.Weir. and Frank R. Giordano. Thomas's Calculus, Pearson 12th Edition 2009.

Unit 4
Descriptive vs. Inferential Statistics

Learning outcome

After the successful completion of the course, the learner will be able to:

- Define descriptive and inferential statistics.
- List components of descriptive statistics (e.g., mean, median).
- Identify common graphs for descriptive statistics.
- Describe purposes of inferential statistics.
- State key differences between descriptive and inferential statistics.

Pre-requisites

Think of descriptive statistics as summarizing information you already have. Imagine you are organizing your favorite books on a shelf. You might count how many books are fiction or non-fiction, find the average number of pages per book, or even note the most common genre. This helps you understand the collection at hand without making any predictions about other books you haven't seen. Descriptive statistics works in the same way—it summarizes data you've already collected to make it easier to understand.

On the other hand, inferential statistics is like making predictions based on your existing collection. Imagine you read the top 10 most popular books from different genres and try to predict the best-selling book of the year based on those. You don't have the data for all books, but you use what you know to make an educated guess about the larger population. Similarly, inferential statistics uses a sample to make generalizations or predictions about a larger group, helping us make decisions even when we don't have all the data.

Keywords

Mean, Median, Mode, Central Tendency, Range, Variance, Standard deviation

Discussion

1.4.1 Descriptive vs. Inferential Statistics

Descriptive statistics refers to the field of methods to organize, display, and describe data through tables, graphs, and summary measures. The purpose of using these statistics is to properly summarize data in an informative way. The basic components that are widely utilized for the purpose are measures of central tendency: mean, median, and mode. Measures of variability include range, variance, or standard deviation. Graphical methods which can be used to display data are histograms, bar charts, pie charts, box plots, or scatter plots. For example, if a teacher records test scores of 30 students, the notion of a mean of the test scores is descriptive of performance summarization. Suppose the student test scores are as follows: [85, 90, 76, 88, 92]. The mean is calculated as (85+90+76+88+92)/5 = 86.2. The use of standard deviation is descriptive in a different example of a survey of 100 households. The variation in monthly electricity bills informs how much these bills differ from the average. In this case, histograms will clearly show the data and how test scores vary or how likely the specific electricity bill is as compared to its mean value.

Inferential statistics, on the other hand, involve techniques that allow us to use samples to make generalizations about a population. The primary purpose is to make inferences or predictions about a population based on sample data. This branch includes sampling methods, hypothesis testing, confidence intervals, and regression analysis. For example, hypothesis testing might be used to determine if a new teaching method is more effective than the traditional one by comparing the test scores of two groups of students. Suppose we compare the mean test scores of students taught by the new method versus the traditional method using a t-test. A **t-test** is a statistical test used to compare the means of two groups and determine if there is a significant difference between them. It is commonly applied when the sample size is small or the population variance is unknown. There are three main types of t-tests:

1. **One-Sample t-test**: Compares a sample mean to a known value.
2. **Independent Two-Sample t-test**: Compares the means of two independent groups.
3. **Paired t-test**: Compares the means of two related groups (e.g., before and after treatment).

Another example is estimating the average height of adult males in a city based on a sample of 100 males. If the sample shows a mean height of 175 cm with a standard deviation of 10 cm, a 95% confidence interval might be calculated to be between 173 cm and 177 cm. Visual representations, such as confidence interval charts, help illustrate the estimated range of values for a population parameter with a certain level of confidence.

The key differences between descriptive and inferential statistics lie in their scope, application, and techniques. Descriptive statistics focus on summarizing the data at hand, while inferential statistics aim to make predictions or inferences about a larger population based on sample data. Descriptive statistics are used to describe the characteristics of a dataset, employing tools like mean, median, mode, standard deviation, and various graphs. In contrast, inferential statistics draw conclusions and make decisions based on data, utilizing techniques such as hypothesis tests, confidence intervals, and regression analysis.

To illustrate, consider a company that surveys 500 customers and finds that the average satisfaction rating is 4.2 out of 5. This finding is an example of descriptive statistics. If the company uses this average satisfaction rating to infer that the average satisfaction of all their customers (beyond the 500 surveyed) is likely around 4.2, and constructs a confidence interval to express the precision of this estimate, they are employing inferential statistics. This comparison highlights how descriptive statistics help us understand and summarize data, while inferential statistics allow us to make predictions and decisions based on data analysis.

1.4.2 Key Concepts: Descriptive Statistics

Descriptive statistics involves methods for organizing, displaying, and describing data. These methods include tables, graphs, and measures that provide insights into the data. The purpose of descriptive statistics is to simplify data presentation, making it easier to understand and interpret.

One part of descriptive statistics is **measures of central tendency**, which include the mean, median, and mode. These measures represent the central or typical values in a dataset. For example, if a teacher records the test scores of 30 students, the mean score shows the overall performance. For scores such as [85, 90, 76, 88, 92], the mean is calculated as (85+90+76+88+92)/5 = 86.2.

Another aspect is **measures of variability**, which include range, variance, standard deviation, and interquartile range. These measures indicate how spread out the data points are. In a survey of 100 households, the standard deviation of monthly electricity bills shows how much the individual bills differ from the average. For bills like [$120, $135, $100, $110, $125], the standard deviation is used to measure this variation.

Graphical tools are also a part of descriptive statistics. These include histograms, bar charts, pie charts, box plots, and scatter plots. They visually represent data, making it easier to identify patterns and relationships. For example, a histogram of test scores for 30 students can show how the scores are distributed.

Descriptive statistics provides useful techniques for understanding and interpreting data through numerical and graphical methods.

1.4.3 Key Concepts: Inferential Statistics

Inferential statistics involves techniques for analyzing sample data to make predictions or draw conclusions about a larger population. These methods help in estimating population characteristics and testing assumptions based on limited data. The purpose of inferential statistics is to make generalizations or predictions about a population using data collected from a representative sample.

Key elements of inferential statistics include **sampling methods**, which are techniques for selecting a subset of data from the population. These methods ensure that the sample accurately represents the population. Another component is **hypothesis testing**, which involves procedures for evaluating assumptions about a population parameter. For example, researchers might test whether a new teaching method is more effective than the traditional one by analyzing the test scores of students taught using each method. A t-test can compare the mean scores of the two groups.

Confidence intervals are another important aspect, providing a range of values to estimate a population parameter. To estimate the average height of adult males in a city, a sample of 100 males with a mean height of 175 cm and a standard deviation of 10 cm might yield a 95% confidence interval of (173 cm, 177 cm). This interval suggests that the true average height of the population lies within this range with 95% certainty.

Additionally, **regression analysis** is used to study relationships between variables, aiding in predictions and understanding dependencies. Graphical tools, such as a confidence interval chart, visually depict estimated ranges for population parameters, making it easier to interpret the data.

Inferential statistics provides tools for drawing conclusions and making informed predictions about populations, using methods that balance statistical rigor and practicality.

1.4.4 Key Differences: Descriptive vs. Inferential Statistics

Descriptive and inferential statistics differ in their purpose, usage, and methods.

1. Descriptive statistics focus on summarizing and presenting the given data. They aim to describe features and patterns within a dataset using methods such as mean, median, mode, and standard deviation. Additionally, visual tools like graphs and charts are employed to effectively represent data and make it easier to interpret.

2. Inferential statistics, on the other hand, aim to predict or infer information about a larger population based on sample data. They are used to draw conclusions and support decision-making by generalizing findings from the sample to the broader group. This approach is particularly valuable when studying populations where direct data collection is impractical.

3. The techniques used in these two branches also differ significantly. Descriptive statistics rely on methods that summarize data, while inferential statistics use tools such as hypothesis testing, confidence intervals, and regression analysis. These methods enable analysts to test theories and make predictions, bridging the gap between data and actionable insights.

4. Inferential statistics utilize methods such as hypothesis testing, confidence intervals, and regression analysis.

Assignment

1. Given the following data set of monthly household expenses: [1200, 1100, 1300, 1250, 1350, 1400, 1100, 1200], find the median and mode. Explain what these measures tell you about the data.
2. Describe different sampling methods (random, stratified, cluster, and systematic) and discuss their advantages and disadvantages. Provide an example of each method.
3. Write an essay comparing and contrasting descriptive and inferential statistics. Include examples of when each type would be appropriately used.

Suggested reading

1. McClave, James T., and Terry Sincich. *Statistics.* 13th ed., Pearson, 2018.
2. Sullivan, Michael. *Fundamentals of Statistics: Informed Decisions Using Data*. 5th ed., Pearson, 2018.
3. Triola, Mario F. *Essentials of Statistics*. 6th ed., Pearson, 2020.
4. Larson, Ron, and Betsy Farber. *Elementary Statistics: Picturing the World*. 7th ed., Pearson, 2020.

Measures of Central Tendency and Dispersion

UNIT-1
Concept of Sampling Distributions; Measure of central tendency

Learning outcome

At the end of this unit, the learner will be able to

- Identify and define the term "sampling distribution" and its purpose in statistics.
- Recall the names of the three measures of central tendency: mean, median, and mode.
- Recognize the formula and method for calculating the mean of a dataset.
- Describe how to find the median in a dataset, including how to handle datasets with an odd or even number of observations.
- Identify the mode of a dataset as the value that appears most frequently

Pre-requisites

Think of a classroom filled with students working on a project. Just like you gather different opinions and results from various students to understand the overall performance of the class, in statistics, we gather data from different samples to understand a broader population.

Sampling Distributions are like collecting feedback from multiple groups of students on the same project. Each group's feedback represents a different sample, and when you combine all these samples, you get an overall picture of how the project was perceived across the entire class. Similarly, in statistics, a sampling distribution helps us understand how the statistics (like means) from various samples can vary and what the overall trend might be.

Measures of Central Tendency—mean, median, and mode—are tools that help us summarize this feedback.

Key Concepts

Sampling Distributions, Central Tendency, Mean, Median, Mode, Variability, Sample Statistics

Discussion

2.1.1 Sampling Distributions

A sampling distribution represents the distribution of a statistic, like the average, derived from numerous random samples taken from a population. It helps us observe how sample statistics, such as sample averages, fluctuate across different samples. By examining the sampling distribution, we can estimate characteristics of the entire population using only a limited number of samples.

Suppose we want to find the average height of students in a large school. Rather than measuring the height of every student, we randomly select several groups of students and calculate the average height for each group. The collection of these sample averages creates what's known as a sampling distribution of the mean.

2.1.2 Introduction to Measures of central tendency

Measures of central tendency, including the mean, median, and mode, are fundamental for summarizing a datasets in a meaningful way. These metrics provide different perspectives on the data, highlighting the central or typical values and giving insights into the distribution and variability of the data. The mean offers an arithmetic average, which is useful for further statistical calculations; the median provides a measure that is less affected by outliers, offering a better central value in skewed distributions; and the mode.

Understanding the concept of sampling distributions and measures of central tendency is crucial in the field of statistics because they provide foundational tools for data analysis and interpretation. Sampling distributions, particularly through the lens of the Central Limit Theorem, allow statisticians to make inferences about population parameters based on sample statistics. This is essential for hypothesis testing, confidence interval construction, and making predictions about larger populations. By understanding how sample means distribute themselves around the population mean, analysts can quantify the reliability and variability of their estimates, leading to more accurate and robust conclusions.

2.1.3 Probability Density Function

A **Probability Density Function (PDF)** is a function that describes the likelihood of a continuous random variable taking on a particular value. Unlike discrete random variables, which have Probability Mass Functions (PMFs), continuous random variables are characterized by PDFs. The PDF gives the relative likelihood for this random variable to take on a given value.

2.1.3.1 Key Properties of a PDF:

1. **Non-Negativity:** The Probability Density Function(PDF) is always non-negative. For a continuous random variable x with

PDF $f(x), \quad f(x) \geq 0 \quad for\ all\ x.$

2. **Normalization:** The total area under the PDF curve is equal to 1, reflecting the fact that the total probability of all possible outcomes is 1. Mathematically,

$$\int_{-\infty}^{\infty} f(x)\, dx = 1.$$

3**. Probability Calculation:** The probability that the random variable x falls within a particular interval $[a, b]$ is given by the integral of the PDF over that interval,

$$P(a \leq X \leq b) = \int_a^b f(x)\, dx.$$

Example:

Consider a continuous random variable x that is normally distributed with a mean μ and standard deviation σ. The PDF of x, denoted as $f(x)$, is given by the formula:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\, exp\, (-\frac{(x-\mu)^2}{2\sigma^2})$$

This function describes the likelihood of x taking on values near the mean μ, with the spread determined by the standard deviation σ. The area under the entire curve of this PDF equals 1, ensuring that the total probability is conserved.

In summary, the PDF is a fundamental concept in probability and statistics, providing a detailed description of the distribution of continuous random variables and enabling the calculation of probabilities over specific intervals.

In the case of discrete random variable, it is called Probability Mass Function (PMF).

2.1.4 Cumulative Distribution Function (CDF)

1. A Cumulative Distribution Function (CDF) describes the probability that a random variable X will take a value less than or equal to x. It provides the cumulative probability up to and including x.

Mathematically, for a random variable X, the CDF is defined as:

$$F(x) = P(X \leq x)$$

Where F(x) is the cumulative distribution function of X

$P(X \leq x)$ is the probability that X takes a value less than or equal to x

2. Key Properties:

- Non-Decreasing: $F(x)$ is non-decreasing, i.e., $F(x_1) \leq F(x_2)\, for\, x_1 \leq x_2$
- Normalization: The CDF ranges from 0 to 1, i.e.,
$$\lim_{x \to -\infty} F(x) = 0 \quad and \quad \lim_{x \to -\infty} F(x) = 1$$
- Probability Calculation: The probability that x falls within the interval $[a, b]$ is given by
$$P(a \leq X \leq b) = F(b) - F(a).$$

3. Mathematical Representation:

- For a continuous random variable x, the CDF $F(x)$ is given by the integral of the PDF from $-\infty\, to\, x$:

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)\, dt$$

4. Example:

- For a normally distributed variable x with mean μ and standard deviation σ, the CDF is:

$$F(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi\sigma^2}} exp(-\frac{(t-\mu)^2}{2\sigma^2})\, dt$$

- This integral does not have a closed-form solution and is typically computed using numerical methods or looked up in standard normal distribution tables.

5. Probability Estimation:

- PDF: Probability over an interval $[a, b]$ is found by integrating the PDF over that interval.

$$P(a \leq X \leq b) = \int_{a}^{b} f(x)\, dx$$

- CDF: Probability over an interval $[a, b]$ is found by subtracting the CDF values at the endpoints.

$$P(a \leq X \leq b) = F(b) - F(a)$$

2.1.5 Introduction to Sampling Distributions

- **Definition:** A sampling distribution is the probability distribution of a given statistic based on a random sample.
- **Purpose:** To understand the behaviour of a statistic over many samples from the same population.
- **Key Concepts:**
  - Population vs. Sample: The population is the entire group we're interested in, while a sample is a subset of the population.
  - Statistic vs. Parameter: A statistic is a measure obtained from a sample (e.g., sample mean), whereas a parameter is a measure obtained from the entire population (e.g., population mean).
  - Central Limit Theorem (CLT): States that the sampling distribution of the sample mean approaches a normal distribution as the sample size increases, regardless of the population's distribution, provided the sample size is sufficiently Large.

2.1.5.1 Characteristics of Sampling Distributions

1. Mean of sampling distribution
   The Mean of sampling distribution of a statistic (eg. sample mean) is equal to the population parameter (eg. population Mean).

$$\mu_{\underline{x}} = \mu$$

   - $\mu_{\underline{x}}$ is the mean of the sampling distribution of sample mean
   - μ is the population mean

2. Spread of sampling distribution (standard error)
   The spread (or variability) of the sampling distribution is described by standard error. It indicates how much the sample statistic (eg. sample mean) varies from sample to sample. For sample mean, the standard error is given by;

$$Standard\ Error\ = \frac{\sigma}{\sqrt{n}}$$

   Where:

   - σ is the population standard deviation
   - n is the sample size

Examples:

Sampling Distribution of the Mean: If we take multiple random samples from a population and calculate the mean for each sample, the distribution of these sample means forms the sampling distribution of the mean.

- Example: Suppose we have a population with a mean (μ) of 50 and a standard deviation (σ) of 10. If we take many samples of size 30, the distribution of the sample means will approximate a normal distribution with a mean of 50 and a standard error of σ/√n (where n is the sample size).

Problems:

1. If the population mean is 100 and the population standard deviation is 15, find the standard error of the mean for samples of size 25.

2. Given a population with a mean of 80 and a standard deviation of 12, calculate the probability that a sample of 40 has a mean greater than 82.

2.1.6 Measures of Central Tendency

- Purpose: To identify the central or typical value in a dataset. This central value gives a quick overview of the data and helps you understand the overall trend or pattern in the dataset.

- Key Measures:

    - Mean: The arithmetic average of a dataset.
    - Median: The middle value when the data is ordered.
    - Mode: The most frequently occurring value in a dataset.

2.1.6.1 Mean

Definition: The sum of all data values divided by the number of values or we can also define as the mean is calculated by adding all the values in a data set together and then dividing the total by the number of values. This gives you the average of the data.

Formula: $Mean\ (\underline{x}) = \frac{\Sigma x_i}{n}$

where Σx_i is the sum of all data values and n is the number of values.

Examples:

Dataset: [10, 20, 30, 40, 50]

Mean = (10 + 20 + 30 + 40 + 50) / 5 = 30

Problems:

1. The number of books read by a student over five months are: 4, 6, 7, 3, and 5. What is the mean number of books read per month?
2. The scores of 4 players in a basketball game are: 18, 22, 15, and 20. What is the mean score of the players?
3. In a survey, the number of hours spent studying by five students in a week were recorded as: 10, 12, 8, 15, and 9 hours. Find the average number of study hours per student.

2.1.6.2 Median

Definition: The median is the middle value in an ordered dataset. If the dataset has an odd number of values, the median is the number that is directly in the center. However, if the dataset has an even number of values, the median is calculated by averaging the two middle numbers.

Finding the Median:

- Order the data from smallest to largest.
- If the number of observations (n) is odd, the median is the middle value.
- If n is even, the median is the average of the two middle values.

Examples:

1. Odd Number of Observations: [5, 10, 15, 20, 25]

   Median = 15

2. Even Number of Observations: [5, 10, 15, 20]

   Median = (10 + 15) / 2 = 12.5

Problems:

1. The following are the ages of 9 students in a class: 18, 21, 25, 23, 19, 20, 22, 24, 26. What is the median age of the students?
2. The number of books read by 6 students in a month are: 3, 5, 7, 4, 6, and 8. Calculate the median number of books read.
3. The weights (in kg) of 7 people are: 55, 60, 65, 70, 75, 80, and 85. What is the median weight?

2.1.6.3 Mode

Definition: The mode is the value that occurs the most frequently in a dataset. If you look at a list of numbers and one number repeats more than the others, that number is the mode. A dataset can have one mode, multiple modes, or no mode if all the numbers appear the same number of times.

Finding the Mode: Identify the value(s) that occur most frequently.

Examples:

1. Single Mode: [4, 5, 6, 6, 7, 8]

   Mode = 6

2. Multiple Modes: [1, 1, 2, 3, 3, 4, 5]

   Modes = 1 and 3

3. No Mode: [1, 2, 3, 4, 5]

   No mode since all values are unique.

Problems:

1. The number of goals scored by a soccer team in five matches are: 2, 3, 3, 1, 4. What is the mode of the goals scored?
2. The following are the shoe sizes of 8 people: 7, 8, 7, 9, 10, 7, 8, 8. Find the mode of the shoe sizes.
3. The number of books read by 6 students are: 5, 3, 5, 4, 5, 2. What is the mode of the books read?

2.1.7 Practice Problems

1. Mean: Given the data [14, 18, 22, 26, 30], calculate the mean.

2. Median: Find the median of the dataset [11, 22, 33, 44, 55, 66, 77, 88].

3. Mode: Determine the mode(s) in the dataset [5, 5, 6, 6, 6, 7, 8, 9, 9].

Mean, median, and mode are fundamental measures of central tendency that provide critical insights into the nature of a dataset. The **mean**, or average, is calculated by summing all the data points and dividing by the number of points, offering a comprehensive measure that incorporates all values. The **Median** is the middle value when the data is ordered, providing a measure that is resistant to outliers and skewed data, thus offering a more robust central value for such distributions. The **mode** is the most frequently occurring value, highlighting the most common data point in the dataset. These measures are crucial in statistics as they summarize large datasets with a single value, making it easier to understand and communicate the general characteristics of the data. They are used extensively in various fields such as economics, psychology, and social sciences to draw meaningful conclusions and support decision-making processes.

Recap

- **Measures of Central Tendency**
  - Mean: arithmetic average, useful for further statistical calculations
  - Median: less affected by outliers, better central value in skewed distributions
  - Mode: most frequently occurring value
- **Importance of Measures of Central Tendency**
  - Provide foundational tools for data analysis and interpretation
  - Enable hypothesis testing, confidence interval construction, and population predictions
- **Sampling Distributions**
  - Probability distribution of a given statistic based on a random sample
  - Central Limit Theorem: sampling distribution of the sample mean approaches a normal distribution as sample size increases
- **Probability Density Function (PDF)**
  - Describes the likelihood of a continuous random variable taking on a particular value
  - Key properties: non-negativity, normalization (area under curve = 1), probability calculation through integration
- **Cumulative Distribution Function (CDF)**
  - Describes the probability that a random variable will take a value less than or equal to a given valu

- Key properties: non-decreasing, normalization (ranges from 0 to 1), probability calculation through differences in CDF values

Objective type questions

1. What is the measure of central tendency that provides the arithmetic average of a dataset?
2. Which measure of central tendency is less affected by outliers?
3. What is the term for the value that appears most frequently in a dataset?
4. What does the Central Limit Theorem state about the sampling distribution of the sample mean as the sample size increases?
5. What is the total area under a Probability Density Function (PDF) curve?
6. What does a Cumulative Distribution Function (CDF) describe about a random variable?
7. If a dataset has an odd number of observations, what is the median value?
8. What is the formula for the Probability Density Function (PDF) of a normally distributed variable with mean μ and standard deviation σ?
9. How is the probability that a random variable falls within a particular interval calculated using the PDF?
10. What does the term "non-negativity" imply about a PDF?
11. What is the key difference between a Probability Density Function (PDF) and a Probability Mass Function (PMF)?
12. What are the key properties of a Cumulative Distribution Function (CDF)?
13. What is the purpose of a sampling distribution?
14. How does the Central Limit Theorem help in hypothesis testing?
15. How is the median calculated if a dataset has an even number of observations?

Answer

1. Mean
2. Median
3. Mode
4. Approaches a normal distribution
5. 1
6. The probability that a random variable will take a value less than or equal to a given value
7. The middle value
8. $\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
9. By integrating the PDF over that interval
10. The PDF is always non-negative
11. PDF is for continuous random variables, PMF is for discrete random variables
12. Non-decreasing, ranges from 0 to 1, probability calculation through differences in CDF values
13. To understand the behavior of a statistic over many samples from the same population
14. By allowing statisticians to make inferences about population parameters based on sample statistics
15. By averaging the two middle values

Assignment

1. Explain the concept of a sampling distribution and its significance in statistical analysis.
2. Calculate the mean, median, and mode of the following dataset: 5, 8, 12, 20, 8, 5, 15.
3. Discuss how the mean, median, and mode can differ in a skewed dataset. Provide examples to illustrate your points.
4. Compare and contrast the properties of the mean, median, and mode as measures of central tendency.
5. Describe how sampling distributions are used to make inferences about a population parameter. Include an example in your explanation.

Suggested Readings

1. Newbold, Paul, William L. Carlson, and Betty Thorne. *Statistics for Business and Economics*. Pearson, 2013.
2. Starnes, Daren S., Dan Yates, and David Moore. *The Practice of Statistics*. W. H. Freeman, 2018.
3. Moore, David S., George P. McCabe, and Bruce A. Craig. *Introduction to the Practice of Statistics*. W. H. Freeman, 2016.
4. Agresti, Alan, and Christine Franklin. *Statistics: The Art and Science of Learning from Data*. Pearson, 2017.

References

1. https://www.geeksforgeeks.org/mean-median-mode/
2. https://nptel.ac.in/courses/110107114

UNIT-2
Range, Interquartile Range (IQR), Concept of Boxplot

Learning outcome

At the end of this unit, the learner will be able to

- Define the range of a dataset.
- Identify the first and third quartiles used in calculating the Interquartile Range (IQR).
- Explain the components of a boxplot.
- Define the interquartile range (IQR).

Pre-requisites

Think about organizing a sports event where you need to compare the performances of different athletes. You might measure their running times or scores. In this situation, you might notice that some athletes perform consistently, while others have varied results. To understand these observations better, we use statistical concepts. These concepts help us summarize and compare data, making it easier to understand.

First, imagine you want to find out the highest and lowest scores among the athletes. This gives you the range, which is the difference between the highest and lowest scores. However, the range alone does not tell you much about the consistency of the performances. To get a clearer picture, we can divide the scores into sections to show how spread out the data is. By breaking the data into quartiles, we can see the spread more clearly. The Interquartile Range (IQR) shows the middle 50% of the data, providing a better measure of consistency than just the range.

To visualize this data, we can use a boxplot, also known as a box-and-whisker plot. This graph shows the range, quartiles, and any outliers in the data set. It gives a quick visual summary of the data distribution, making it easier to compare different groups. Learning these concepts will help you summarize and analyze data more effectively, similar to how you would organize and compare athletes' performances in a sports event.

Key Concepts

Range, Interquartile Range (IQR), Boxplot, Quartiles, Outliers

Discussion

2.2.1 Overview of Range, Interquartile Range (IQR), Concept of Boxplot

In data analysis, understanding the spread and distribution of data is crucial for meaningful interpretation. Three key concepts that help in summarizing and analyzing data are Range, Interquartile Range (IQR), and Boxplot. Each of these tools provides insights into different aspects of data variability and distribution.

Range is the simplest measure of data spread, calculated by subtracting the smallest value from the largest value in a dataset. It provides a quick sense of the overall span of the data but can be influenced by extreme values or outliers. While the range gives a basic idea of data dispersion, it may not fully capture the distribution's central tendencies or the concentration of values.

Interquartile Range (IQR) offers a more refined measure of variability. By focusing on the middle 50% of the data, the IQR is calculated as the difference between the first quartile (Q1) and the third quartile (Q3). This measure effectively filters out the effects of extreme values, providing a clearer picture of the data's central distribution and consistency.

A **Boxplot** (or Box-and-Whisker Plot) visually represents these concepts, offering a graphical summary of data distribution. It displays the median, quartiles, and range of the data, along with potential outliers. The box in the plot shows the IQR, while the "whiskers" extend to the minimum and maximum values within a specified range. This visualization helps quickly identify the spread, central tendency, and any irregularities in the data, making it an essential tool for data analysis and comparison.

2.2.2 Range and the Interquartile Range

In statistics, measures of spread are used to describe the variability or dispersion within a dataset. Two commonly used measures are the Range and the Interquartile Range (IQR). The Range is the simplest measure, calculated as the difference between the highest and lowest values in a dataset. For example, in the dataset $4, 8, 15, 16, 23, 42$, the Range is 38, calculated by subtracting the minimum value (4) from the maximum value (42).

2.2.2.1 Quartiles

Quartiles are values that divide a dataset into four equal parts, each containing a quarter of the data. They are used to describe the spread and centre of a data distribution and provide insight into the distribution's shape and variability.

There are three key quartiles in any dataset:

1. **First Quartile (Q1)**: Also known as the lower quartile, Q1 is the median of the lower half of the dataset. It separates the lowest 25% of the data from the rest.
2. **Second Quartile (Q2)**: Also known as the median, Q2 divides the dataset in half, separating the lower 50% from the upper 50%.
3. **Third Quartile (Q3)**: Also known as the upper quartile, Q3 is the median of the upper half of the dataset. It separates the highest 25% of the data from the rest.

How to Calculate Quartiles

1. Arrange the Data: Sort the dataset in ascending order.
2. Find the Median (Q2): The median is the middle value of the dataset. If the dataset has an odd number of observations, the median is the middle number. If the dataset has an even number of observations, the median is the average of the two middle numbers.
3. Find Q1: Q1 is the median of the lower half of the data (not including the overall median if the number of observations is odd).
4. Find Q3: Q3 is the median of the upper half of the data (not including the overall median if the number of observations is odd).

Example

Consider the dataset: $4, 8, 15, 16, 23, 42$

1. Arrange the data: $4, 8, 15, 16, 23, 42$

2. Find the Median (Q2): The dataset has 6 observations. The median is the average of the third and fourth values.

$$Q2 = \frac{15+16}{2} = 15.5$$

3. Find Q1: The lower half of the data is $4, 8, 15$. The median of this subset is 8.

$$Q1 = 8$$

4. Find Q3: The upper half of the data is [16, 23, 42]. The median of this subset is 23.

$$Q3 = 23$$

2.2.3 Importance of Quartiles

Quartiles provide a way to summarize the distribution of a dataset by identifying key points that divide the data into quarters. They are especially useful for:

- Understanding Data Spread: Quartiles help identify the range within which the central half of the data falls.
- Detecting Outliers: Values significantly below Q1 or above Q3 can be considered outliers.
- Comparing Distributions: Quartiles allow for easy comparison of the spread and centre of different datasets.

By dividing data into four equal parts, quartiles offer a simple yet powerful tool for data analysis and interpretation.

The Interquartile Range (IQR) is a more robust measure of dispersion, as it is not affected by extreme values or outliers. It measures the spread of the middle 50% of the data. To calculate the IQR, one must first arrange the data in ascending order and then find the first quartile (Q1), which is the median of the lower half, and the third quartile (Q3), which is the median of the upper half. The IQR is the difference between Q3 and Q1. For example, in the dataset

$4, 8, 15, 16, 23, 42$, the IQR is 15, calculated by subtracting Q1 (8) from Q3 (23).

These measures of spread help describe the variability within a dataset. While the Range provides a quick measure of overall spread, the IQR offers a more robust measure by focusing on the central portion of the data, thus reducing the impact of outliers. Understanding these concepts is crucial for analysing and interpreting data effectively.

2.2.4 Boxplot

A boxplot, also known as a box-and-whisker plot, is a graphical representation of a dataset that shows its distribution, central values, and variability. It provides a visual summary of several key statistical measures: the minimum, first quartile (Q1), median (Q2), third quartile (Q3), and maximum.

Components of a Boxplot:

1. **Box**: The central part of the plot, which extends from Q1 to Q3, representing the interquartile range (IQR). This box captures the middle 50% of the data.
2. **Median Line**: A line inside the box that shows the median (Q2) of the data, which divides the dataset in half.
3. **Whiskers**: Lines that extend from the edges of the box to the minimum and maximum values within a certain range (often 1.5 times the IQR). These lines show the spread of the majority of the data.
4. **Outliers**: Individual points plotted outside the whiskers, representing data points that fall significantly below Q1 or above Q3. These outliers can indicate variability in the data or potential errors.

2.2.4.1 Why should one use a Boxplot?

Boxplots are incredibly useful for understanding the distribution of data at a glance. They help identify:

- **Central Value**: The median line inside the box shows the centre of the data.
- **Spread of the Data**: The length of the box (IQR) and the whiskers show how spread out the data points are.
- **Symmetry or Skewness**: The relative position of the median line within the box can indicate whether the data is symmetric or skewed.
- **Outliers**: Points outside the whiskers highlight potential outliers that might need further investigation.

Example:

Imagine you have test scores from a class:55, 60, 67, 70, 72, 75, 78, 80, 82, 85, 87, 90, 92, 95,98.

To create a boxplot:

1. **Arrange the data**: 55, 60, 67, 70, 72, 75, 78, 80, 82, 85, 87, 90, 92, 95, 98
2. **Calculate the Quartiles**:
    - Q1 (25th percentile) = 71.0
    - Median (Q2, 50th percentile) = 80
    - Q3 (75th percentile) = 88.5
3. **Identify the Minimum and Maximum**:
    - Minimum = 55
    - Maximum = 98
4. **Draw the Boxplot**: The box extends from 71 (Q1) to 88.5 (Q3), with a line at 80 (median). Whiskers extend from 55 (minimum) to 98 (maximum) as shown in figure 2.2.1. Any points beyond these whiskers would be considered outliers.

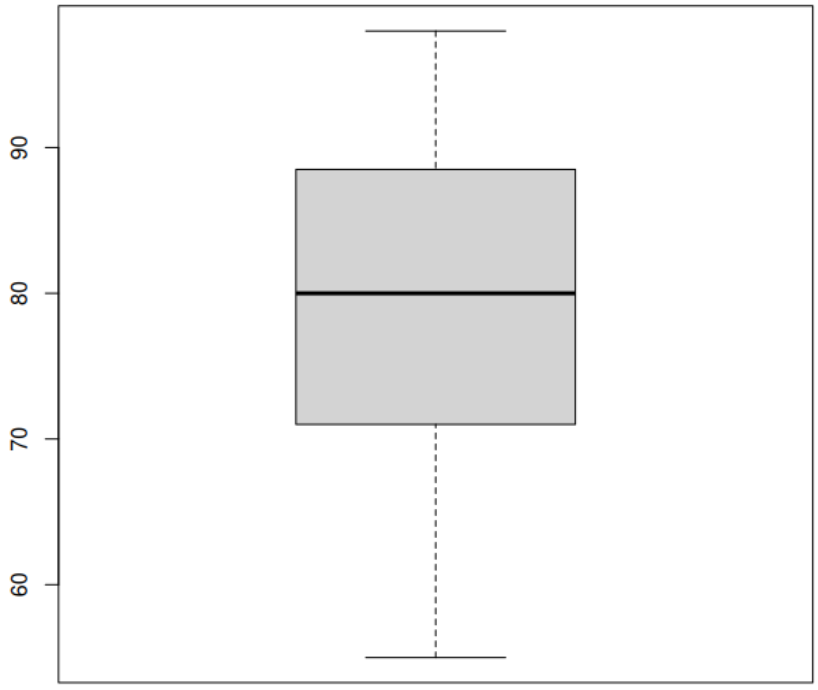


Figure 2.2.1 Boxplot

Boxplots provide a concise summary of the data distribution, making them a valuable tool for exploratory data analysis. They allow you to quickly compare different datasets and identify patterns, trends, and outliers with ease.

2.2.4.2 Sample Questions

1. The heights (in cm) of students in a class are given below:

   150, 155, 160, 165, 170, 175, 180, 185, 190, 195, 200, 150, 155, 160, 165, 170, 175, 180, 185, 190, 195, 200

- Determine the five-number summary for the data.
- Compute the interquartile range (IQR).
- Check for any outliers in the dataset.
- Create a boxplot to visually represent the data.

2. The marks scored by 15 students in a mathematics test are:

   35, 40, 45, 50, 55, 60, 65, 70, 75, 80, 85, 90, 95, 100, 105

   - Compute the minimum, first quartile (Q1), median, third quartile (Q3), and maximum.
   - Calculate the IQR and identify any outliers.
   - Plot the boxplot.

2.2.5 Comparison of Range, Interquartile Range (IQR), Boxplot

Table 2.2.1 comparison of Range, Interquartile Range and Boxplot

Criteria	Range	Interquartile Range (IQR)	Boxplot
Definition	Difference between maximum and minimum values.	Spread of the middle 50% of the data (Q3 - Q1).	Graphical representation showing the distribution of a dataset.
Use	Measures total data spread	Measures spread while ignoring outliers.	Visualizes central tendency, spread, and outliers.
Calculation	$Range = Max - Min$	$IQR = Q3 - Q1$	Plots based on key values: minimum, first quartile ($Q1$), median, third quartile ($Q3$), and maximum.

Recap

➢ **Range**:

- Difference between the maximum and minimum values.

- $Range = Max - Min$
- Provides a basic measure of data spread.
- Can be influenced by outliers.

➢ **Interquartile Range (IQR)**:

- Difference between the third quartile (Q3) and the first quartile (Q1).
- $IQR = Q3 - Q1$
- Represents the spread of the middle 50% of the data.
- Less affected by outliers compared to the range.

➢ **Boxplot**:

- Visual representation showing median, quartiles, and range.
- **Box**: Extends from Q1 to Q3, representing the IQR.
- **Whiskers**: Extend from the box to the minimum and maximum values within a specified range.
- **Outliers**: Points outside the whiskers, indicating variability or potential anomalies.

Objective type questions

1. Define the Range of a dataset and explain how it is calculated.
2. What is the Interquartile Range (IQR)?
3. List the three quartiles in a dataset
4. How do you calculate the first quartile (Q1) and the third quartile (Q3) in a dataset?
5. What information does the Range provide about a dataset?
6. Explain why the Interquartile Range (IQR) is considered a robust measure of dispersion.
7. In a boxplot, what do the whiskers represent, and how are they determined?
8. How is the median (Q2) represented in a boxplot, and what does it indicate about the dataset?
9. Describe the process for creating a boxplot from a given dataset.
10. What are outliers in a boxplot?
11. If the Range of a dataset is 30 and the Interquartile Range (IQR) is 10, what can be inferred about the distribution of the data?
12. How do you find the median value of a dataset with an even number of observations?
13. What role do quartiles play in understanding the spread and center of a dataset?
14. Why might you prefer using the IQR over the Range when analyzing the spread of data?
15. Explain how quartiles can be used to compare different datasets.

Answer

1. **Range**: Difference between the highest and lowest values in a dataset.

i. Range = Max - Min

2. **Interquartile Range (IQR)**: Difference between the third quartile (Q3) and the first quartile (Q1).
3. **Quartiles**:

a. **Q1**: Median of the lower half.

b. **Q2**: Median of the dataset.

c. **Q3**: Median of the upper half.

4. **Calculating Quartiles**: Sort data, find the median of the lower and upper halves.
5. **Range**: Shows the spread from the lowest to the highest value.
6. **IQR**: Measures the spread of the middle 50% of data, reducing outlier impact.
7. **Whiskers**: Extend from Q1 to minimum and Q3 to maximum, within 1.5 times the IQR.
8. **Median (Q2)**: Line inside the box in a boxplot, dividing the dataset in half.
9. **Boxplot Creation**: Arrange data, calculate quartiles, draw box from Q1 to Q3, mark Q2, extend whiskers, plot outliers.
10. **Outliers**: Data points outside the whiskers of a boxplot.
11. **Range vs. IQR**: Range shows overall spread; IQR shows spread of the middle 50% of data.
12. **Median for Even Observations**: Average of the two middle values.
13. **Quartiles**: Divide data into four parts to understand distribution and variability.
14. **IQR vs. Range**: IQR is preferred for robustness against outliers.
15. **Comparing Quartiles**: Helps in analyzing data spread and distribution across datasets.

Assignment

1. Given the dataset [5, 7, 8, 12, 15, 20, 25, 30, 35, 40], calculate the Range and Interquartile Range (IQR). Explain the significance of each measure in describing the dataset's spread.
2. For the dataset [10, 12, 15, 18, 20, 22, 25, 30, 35, 40], determine the first quartile (Q1), median (Q2), and third quartile (Q3). Use these values to calculate the Interquartile Range (IQR).
3. Create a boxplot for the dataset [3, 7, 8, 12, 14, 18, 20, 22, 24, 30]. Include calculations for the quartiles, median, and whiskers. Identify and explain any potential outliers.
4. Compare the spread of two datasets using their Range and IQR. Given the datasets [2, 4, 6, 8, 10, 12, 14] and [5, 7, 9, 11, 13, 15, 17, 19], calculate the Range and IQR for each and discuss which dataset has greater variability and why.
5. Analyze a boxplot representing the following statistical summary: Minimum = 5, Q1 = 10, Median (Q2) = 15, Q3 = 20, Maximum = 25. Describe the spread of the data, the position of the median, and the presence of any outliers based on the boxplot information.

Suggested Readings

1. Mann, Peggy. *Introduction to Statistics*. 5th ed., Cengage Learning, 2017.
2. Agresti, Alan, and Barbara Finlay. *Statistical Methods for the Social Sciences*. 5th ed., Pearson, 2018.
3. Moore, David S., George P. McCabe, and Bruce A. Craig. *Introduction to the Practice of Statistics*. 9th ed., W.H. Freeman, 2021.
4. Johnson, Richard A., and Dean W. Wichern. *Applied Multivariate Statistical Analysis*. 6th ed., Pearson, 2014.
5. Sharma, Sanjay. *The Essence of Multivariate Thinking: Basic Themes and Methods*. Routledge, 1996.

References

1. https://www.geeksforgeeks.org/mean-median-mode/
2. https://nptel.ac.in/courses/110107114

UNIT-3

Variance and Standard Deviation; Population and Sample variance

Learning outcome

- Define variance and standard deviation in the context of statistical data analysis.
- Identify the differences between population variance and sample variance.
- Calculate the variance and standard deviation for a given dataset.
- List the formulas used to compute population variance, sample variance, and standard deviation.

Pre-requisites

Imagine you have a basket of apples with varying sizes. If you want to know how much the sizes of the apples differ from each other, you could compare each apple to the average size. This is similar to how we measure the spread or dispersion in statistics. Just as you'd measure how much each apple's size deviates from the average, variance and standard deviation measure how individual data points in a dataset differ from the mean.

Variance is like calculating the average of these size differences squared. It gives you a sense of how much the sizes vary on average. **Standard deviation** is the square root of the variance, which brings the measure back to the same units as the original data (like the size of the apples), making it easier to interpret.

For **population variance** and **sample variance**, think of the difference between having a full basket of apples versus just a small sample of them. Population variance considers the variability within the entire basket, while sample variance estimates the variability based on just a few apples. Understanding these concepts helps in analyzing and summarizing data effectively, similar to assessing the consistency of apple sizes in different scenarios.

Key Concepts

Variance, Standard Deviation, Population Variance, Sample Variance, Dispersion

Discussion

2.3.1 Variance and Standard Deviation

In statistics, understanding the spread or variability of a dataset is crucial. Two fundamental measures of this variability are Variance and Standard Deviation. These metrics provide insights into how much the data points differ from the mean (average) of the dataset. Variance calculates the average squared difference between each data point and the mean, offering a general view of the data's spread. However, since variance is expressed in squared units, it can be difficult to interpret directly. This is where the standard deviation becomes more intuitive, as it is the square root of the variance, restoring the original units of the data and making it more comprehensible. Both variance and standard deviation are critical in statistical analysis, providing valuable insights into how consistent or dispersed the data is.

2.3.1.1 Variance

Variance measures the average squared deviation of each data point from the mean. It gives us an idea of how spread out the data points are in a dataset.

Formula for Variance:

Population Variance (σ^2): Used when you have data for the entire population.

$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$ Were

N : Number of data points in the population

x_i: Each data point and

μ: Mean of the population

Sample Variance (s^2): Used when you have a sample from the population.

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \underline{x})^2}{n-1}$$

were

n : Number of data points in the sample

x_i : Each data point

$\underline{x}$: Mean of the sample

The key difference is that sample variance uses $n-1$ (degrees of freedom) instead of n to provide an unbiased estimate of the population variance.

2.3.2 Degrees of Freedom in Statistics

Degrees of freedom is a concept in statistics that refers to the number of independent values or quantities that can vary in the analysis without breaking any constraints. It is an essential idea in statistical calculations and helps in understanding the variability in data and the reliability of statistical estimates.

When calculating statistics such as variance, standard deviation, and certain hypothesis tests, degrees of freedom are used to account for the number of values that are free to vary. For instance, when estimating the sample variance, we use $n-1$ (where n is the sample size) as the degrees of freedom. This adjustment is necessary because one value is constrained by the requirement that the sum of deviations from the mean equals zero.

Why is degrees of freedom important?

Degrees of freedom play a crucial role in:

1. Estimating Parameters: They help in estimating population parameters more accurately.
2. Hypothesis Testing: They are used in various test statistics, such as the t-test, chi-square test, and F-test, to determine critical values and p-values.
3. Adjusting for Bias: When calculating sample statistics, degrees of freedom help adjust for the bias that arises from using sample data instead of the entire population.

Example:

Consider you have a sample of five test scores: $80, 85, 90, 95, 100$

1. Calculate the Mean ($\underline{x}$):

$$\underline{x} = \frac{80+85+90+95+100}{5} = 90$$

2. Compute Deviations from the Mean:

$$(80-90), (85-90), (90-90), (95-90), (100-90) = -10, -5, 0, 5, 10$$

3. Sum of Deviations:

The sum of deviations from the mean is always zero:

$$(-10) + (-5) + 0 + 5 + 10 = 0$$

4. Degrees of Freedom:

When calculating the sample variance, we use $n-1 = 5-1 = 4$ degrees of freedom. This is because once the mean is fixed, only four of the five values can vary independently; the fifth value is constrained by the requirement that the sum of deviations equals zero.

2.3.3 Standard Deviation

Standard Deviation is the square root of the variance. It is a more intuitive measure of spread because it is in the same units as the data points.

Formula for Standard Deviation:

1. Population Standard Deviation (σ):

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}}$$

2. Sample Standard Deviation (s):

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \underline{x})^2}{n-1}}$$

Example

Population Variance and Standard Deviation

Consider a small population of test scores: $5, 7, 8, 9, 10$

1. Calculate the Mean (μ):

$$\mu = \frac{5+7+8+9+10}{5} = 7.8$$

2. Calculate each squared deviation from the mean and sum them:

$$\Sigma(x_i - \mu)^2 = (5 - 7.8)^2 + (7 - 7.8)^2 + (8 - 7.8)^2 + (9 - 7.8)^2 + (10 - 7.8)^2 = 14.8$$

3. Calculate the Population Variance (σ^2):

$$\sigma^2 = \frac{14.8}{5} = 2.96$$

4. Calculate the Population Standard Deviation (σ):

$$\sigma = \sqrt{2.96} \approx 1.72$$

2.3.4 Sample Variance and Standard Deviation

Consider a sample of test scores: $5, 7, 8, 9, 10$ (same as above, but treat as a sample)

1. Calculate the Mean ($\underline{x}$):

$$\underline{x} = \frac{5+7+8+9+10}{5} = 7.8$$

2. Calculate each squared deviation from the mean and sum them:

$$\Sigma(x_i - \underline{x})^2 = 14.8 \text{ (same as above)}$$

3. Calculate the Sample Variance (s^2):

$$s^2 = \frac{14.8}{5-1} = \frac{14.8}{4} = 3.7$$

4. Calculate the Sample Standard Deviation (s):

$$s = \sqrt{3.7} \approx 1.92$$

Variance and Standard Deviation are essential tools for understanding the spread of data. The Population Variance and Standard Deviation provide measures for the entire population, while the Sample Variance and Standard Deviation offer estimates based on a sample. These metrics help us understand how much individual data points deviate from the mean, providing valuable insights into the dataset's variability.

Recap

- **Variance**: Measures the average squared deviation of each data point from the mean, indicating data spread.
- **Population Variance (σ^2)**: Used for entire population data.
  - Formula: $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$
- **Sample Variance (s^2)**: Used for sample data, providing an unbiased estimate of population variance.
  - Formula: $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$
- **Difference**: Sample variance uses n−1 (degrees of freedom) instead of n.
- **Degrees of Freedom**: Refers to the number of independent values in a calculation.
- Important for accurate parameter estimation and hypothesis testing.
- **Example**: In a sample variance calculation, degrees of freedom is n−1, reflecting the constraint that deviations from the mean must sum to zero.
- **Standard Deviation**: The square root of variance, providing a measure of spread in the same units as the data.
- **Population Standard Deviation (σ)**:
  - Formula: $\sigma = \sqrt{\sigma^2}$
- **Sample Standard Deviation (s)**:
  - Formula: $s = \sqrt{s^2}$

Objective type questions

1. Define Variance and explain its purpose in statistics.
2. Calculate the Population Variance given the following data: 10, 12, 14, 16, 18.
3. What is the formula for Sample Variance?
4. Explain the concept of Degrees of Freedom in the context of calculating Sample Variance.
5. Compute the Sample Standard Deviation for the following dataset: 4, 8, 6, 5, 7.
6. Describe the relationship between Standard Deviation and Variance.
7. How do you adjust the formula for Variance when dealing with a sample rather than a population?
8. Calculate the Population Standard Deviation for the dataset: 20, 22, 24, 26, 28.
9. Explain why the Degrees of Freedom are used when calculating Sample Variance.
10. What does a larger Standard Deviation indicate about a dataset?
11. Calculate the Variance of a dataset consisting of the values: 5, 10, 15, 20, 25.
12. What is the primary difference between Population Standard Deviation and Sample Standard Deviation?
13. Given a dataset of 12, 15, 20, 22, and 28, find the Sample Variance.
14. Explain how Variance and Standard Deviation are useful for understanding the spread of data.
15. Calculate the Standard Deviation for a sample dataset with the values: 3, 7, 5, 9, 6.

Answer

1. Variance measures the average squared deviation of each data point from the mean, indicating how spread out the data points are.
2. Population Variance = 8.0
3. Sample Variance = $\Sigma (xi - \bar{x})^2 / (n - 1)$
4. Degrees of Freedom represent the number of independent values that can vary. For Sample Variance, it is (n - 1) to correct for the bias in estimating the population variance.
5. Sample Standard Deviation ≈ 1.41
6. Standard Deviation is the square root of Variance. It provides a measure of spread in the same units as the data.
7. The formula for Sample Variance uses (n - 1) in the denominator, unlike Population Variance which uses n, to account for the estimation error.
8. Population Standard Deviation ≈ 3.16
9. Degrees of Freedom are used to adjust for the fact that one value is constrained by the mean, leading to (n - 1) in the denominator.
10. A larger Standard Deviation indicates greater variability or spread in the dataset.
11. Variance = 62.5
12. Population Standard Deviation is for the entire population, while Sample Standard Deviation is an estimate based on a sample and uses (n - 1) for unbiased estimation.
13. Sample Variance ≈ 41.67
14. Variance and Standard Deviation provide insights into the spread and dispersion of data, helping to understand how data points differ from the mean.
15. Sample Standard Deviation ≈ 2.55

Assignment

1. Calculate the variance and standard deviation for the dataset: [8, 12, 14, 10, 18].
2. Given the sample data: [5, 7, 8, 6, 9], find the sample variance and standard deviation.
3. Explain why we use (n - 1) in the calculation of sample variance.
4. For the following numbers: [2, 4, 6, 8, 10], compute the range, variance, and standard deviation.
5. A sample of exam scores is: [75, 80, 85, 90, 95]. Find the sample variance and standard deviation.

Suggested Readings

1. Mann, Peggy. *Introduction to Statistics*. 5th ed., Cengage Learning, 2017.
2. Agresti, Alan, and Barbara Finlay. *Statistical Methods for the Social Sciences*. 5th ed., Pearson, 2018.
3. Moore, David S., George P. McCabe, and Bruce A. Craig. *Introduction to the Practice of Statistics*. 9th ed., W.H. Freeman, 2021.
4. Johnson, Richard A., and Dean W. Wichern. *Applied Multivariate Statistical Analysis*. 6th ed., Pearson, 2014.
5. Sharma, Sanjay. *The Essence of Multivariate Thinking: Basic Themes and Methods*. Routledge, 1996.

References

1. https://www.geeksforgeeks.org/mean-median-mode/
2. https://nptel.ac.in/courses/110107114

UNIT-4
Sampling Distribution of the Sample Mean and Sample Proportion

Learning outcome

At the end of this unit, the learner will be able to;

- Identify the formula for calculating the Coefficient of Variation
- Recall the concept of the sampling distribution of the sample mean
- Recognize the characteristics of the sampling distribution of the sample proportion
- List different types of graphs used to summarize data involving multiple variables

Pre-requisites

To understand the Coefficient of Variation, Sampling Distribution, and summarizing data with multiple variables, let's build on familiar concepts. The Coefficient of Variation helps us gauge the relative variability of data by comparing the standard deviation to the mean. This is like comparing the consistency of different products based on their performance relative to their average.

When discussing the Sampling Distribution of the Sample Mean, think of it as estimating the average performance of a product based on several small samples rather than testing every item. It gives you insights into the reliability of your estimates for the entire population.

Finally, summarizing data from multiple variables using graphs and correlation allows you to visualize and understand complex relationships between different aspects of your data. This is similar to creating a comprehensive report where various metrics are analyzed together to give a complete picture of performance. By exploring these topics, you will enhance your ability to interpret and analyze diverse datasets effectively.

Key Concepts

Coefficient of Variation, Sampling Distribution, Standard Deviation, Correlation, Scatter Plot

Discussion

2.4.1 Coefficient of Variation

When comparing the variability of two or more datasets, especially those with different units or widely different means, the Coefficient of Variation (CV) becomes a valuable tool. The CV is a standardized measure of dispersion of a probability distribution or frequency distribution. It is often expressed as a percentage and is used to compare the degree of variation between datasets.

The Coefficient of Variation is defined as the ratio of the standard deviation to the mean. It provides a relative measure of dispersion in relation to the mean, making it dimensionless.

Formula:

$$CV = (\frac{\sigma}{\mu})X\ 100\%$$

were

σ= Standard deviation of the dataset

μ= Mean of the dataset

2.4.1.1 Why should one use the Coefficient of Variation?

1. Comparing Different Datasets: CV is useful when you need to compare the variability of datasets with different units or scales. For example, comparing the variation in heights (measured in centimetres) and weights (measured in kilograms) of a group of people.
2. Understanding Relative Risk: In finance, the CV is used to assess the risk per unit of return. A lower CV indicates a more stable investment.
3. Normalizing Data: By expressing variability relative to the mean, the CV helps normalize data, making it easier to compare.

Example

Consider two datasets representing the test scores of two different classes:

Class A: [70, 75, 80, 85, 90]

Class B: [60, 70, 80, 90, 100]

Step-by-Step Calculation for Class A:

1. Calculate the Mean (μ):

$$\mu = \frac{70+75+80+85+90}{5} = 80$$

2. Calculate the Standard Deviation (σ):

$$\sigma = \sqrt{\frac{(70-80)^2+(75-80)^2+(80-80)^2+(85-80)^2+(90-80)^2}{5}}$$

$$= \frac{\sqrt{100+25+0+25+100}}{5} = \sqrt{50} \approx 7.07$$

3. Calculate the Coefficient of Variation (CV):

$$CV = (\frac{7.07}{80})\text{X}100\% \approx 8.84\%$$

Step-by-Step Calculation for Class B:

1. Calculate the Mean (μ):

$$\mu = \frac{60+70+80+90+100}{5} = 80$$

2. Calculate the Standard Deviation (σ):

$$\sigma = \sqrt{\frac{(60-80)^2+(70-80)^2+(80-80)^2+(90-80)^2+(100-80)^2}{5}}$$

$$= \sqrt{\frac{400+100+0+100+400}{5}} = \sqrt{200} \approx 14.14$$

3. Calculate the Coefficient of Variation (CV):

$$CV = (\frac{14.14}{80})X100\% \approx 17.68\%$$

2.4.2 Interpretation

In this example, Class A has a CV of approximately 8.84%, while Class B has a CV of approximately 17.68%. This indicates that the scores in Class A are more consistent relative to their mean compared to those in Class B. Even though both classes have the same mean score, the higher CV in Class B signifies greater relative variability.

The Coefficient of Variation is a powerful tool for comparing the relative variability of different datasets. By standardizing the measure of dispersion, it allows for meaningful comparisons across different units and scales. Whether in finance, research, or everyday data analysis, understanding and applying the CV can provide deeper insights into data behaviour and variability.

2.4.3 Sampling Distribution of the Sample Mean and Sample Proportion

In statistics, understanding the behaviour of sample statistics (such as the sample mean and sample proportion) is crucial for making inferences about the population. The sampling distribution describes how these statistics vary from sample to sample. Two key types of sampling distributions are the sampling distribution of the sample mean and the sampling distribution of the sample proportion.

2.4.3.1 Sampling Distribution of the Sample Mean

The sampling distribution of the sample mean describes how the mean of a sample ($\underline{x}$) varies from sample to sample. It's particularly important because it allows us to make inferences about the population mean (μ).

Properties:

1. Mean: The mean of the sampling distribution of the sample mean is equal to the population

mean (μ). That is E ($\underline{x}$) = μ

2. Standard Deviation: The standard deviation of the sampling distribution (also called the standard error) is equal to the population standard deviation (σ) divided by the square root of the sample size (n).

That is, $\sigma_{\underline{x}} = \frac{\sigma}{\sqrt{n}}$

3. Shape: If the population distribution is normal, the sampling distribution of the sample mean will also be normal. According to the Central Limit Theorem, even if the population distribution is not normal, the sampling distribution of the sample mean will approach a normal distribution as the sample size increases.

Example:

Suppose we have a population with a mean (μ) of 50 and a standard deviation (σ) of 10. We take a sample of size $n = 25$.

1. Mean of the sampling distribution:

$$E(\underline{x}) = 50$$

2. Standard error:

$$\sigma_{\underline{x}} = \frac{10}{\sqrt{25}} = 2$$

The sampling distribution of the sample mean will have a mean of 50 and standard error of 2. This is illustrated in the following figure 2.4.1. Let take 1000 sample sets with size 25. For each sample set we can find the mean. So, we will get 1000 numbers and each such number is the mean of a sample set. Then we can divide the range of the set of means in several bins. The vertical bar in the following figure 2.4.1 represent the number of mean numbers in that bin divided by 1000 which is probability of a mean number to fall in the that bin. The continuous curve in the figure 2.4.1, the population distribution. The vertical bars correspond to the sampling distribution. As can be seen from the figure 2.4.1 that mean of the sampling distribution is the population distribution.

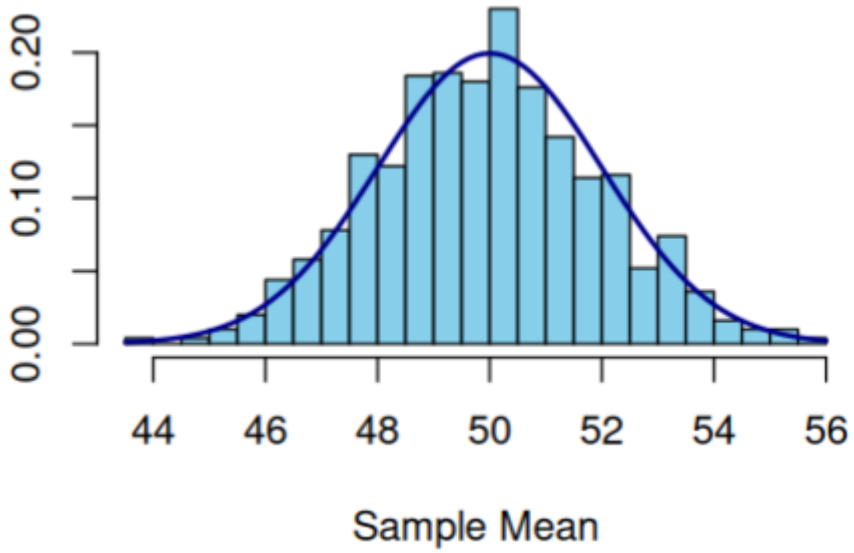


Figure 2.4.1 sampling distribution of the sample mean

2.4.4 Sampling Distribution of the Sample Proportion

The sampling distribution of the sample proportion describes how the proportion of a sample ($\hat{p}$) varies from sample to sample. It is used to make inferences about the population proportion (p).

Properties:

1. Mean: The mean of the sampling distribution of the sample proportion is equal to the population proportion (p).

$$E(\hat{p}) = p$$

2. Standard Deviation: The standard deviation of the sampling distribution (standard error) is given by:

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

3. Shape: For large sample sizes, the sampling distribution of the sample proportion will be approximately normal.

Example:

Suppose we have a population proportion (p) of 0.6 and we take a sample of size $n = 100$.

1. Mean of the sampling distribution: $E(\hat{p})$=0.6

2. Standard error:

$$\sigma_{\hat{p}} = \sqrt{\frac{0.6\ X\ 0.4}{100}} = \sqrt{0.0024} \approx 0.049$$

The sampling distribution of the sample proportion will have a mean of 0.6 and a standard error of 0.049.

Problems and Solutions

Problem 1: Sample Mean

A population has a mean of 100 and a standard deviation of 20. A sample of size 64 is taken. Find the mean and standard error of the sampling distribution of the sample mean.

Solution:

1. Mean of the sampling distribution:

$E(\underline{x})$=100

since the mean of the sample means is same the mean of the population.

2. Standard error:

$$\sigma_{\underline{x}} = \frac{20}{\sqrt{64}} = \frac{20}{8} = 2.5$$

The sampling distribution of the sample mean will have a mean of 100 and a standard error of 2.5.

Problem 2: Sample Proportion

A population proportion is 0.7. A sample of size 200 is taken. Find the mean and standard error of the sampling distribution of the sample proportion.

Solution:

1. Mean of the sampling distribution:

$E(\hat{p})$=0.7

2. Standard error:

$$\sigma_{\hat{p}} = \sqrt{\frac{0.7\ X\ 0.3}{200}} = \sqrt{0.00105} \approx 0.032$$

The sampling distribution of the sample proportion will have a mean of 0.7 and a standard error of 0.032.

Understanding the sampling distributions of the sample mean and sample proportion is crucial for making statistical inferences about populations. These distributions provide insights into the variability of sample statistics and help determine the precision of sample estimates. By understanding these concepts, we can make more informed decisions and draw more accurate conclusions from our data.

2.4.5 Summarizing Data from More Than One Variable

When working with data, it's often essential to look at relationships between multiple variables. This can help us understand how variables interact with each other. Two key tools for this are graphs and correlation analysis. In this lecture, we'll explore different ways to visualize relationships between variables and how to measure their strength using correlation. We'll also work through some examples and problems.

Correlation is a statistical measure that describes the strength and direction of the relationship between two variables. Understanding correlation is crucial for identifying relationships and making predictions in various fields, from finance to biology.

2.4.6 What is Correlation?

Correlation quantifies the degree to which two variables move in relation to each other. It indicates how changes in one variable are associated with changes in another variable. The most common correlation coefficient is the Pearson correlation coefficient, denoted by r.

2.4.6.1 Pearson Correlation Coefficient

The Pearson correlation coefficient ranges from -1 to 1:

r = +1 indicates a perfect positive linear relationship, meaning that as one variable increases, the other also increases.

r = -1 indicates a perfect negative linear relationship, meaning that as one variable decreases, the other also decreases.

r = 0 indicates no linear relationship, meaning that no relationship between the variables.

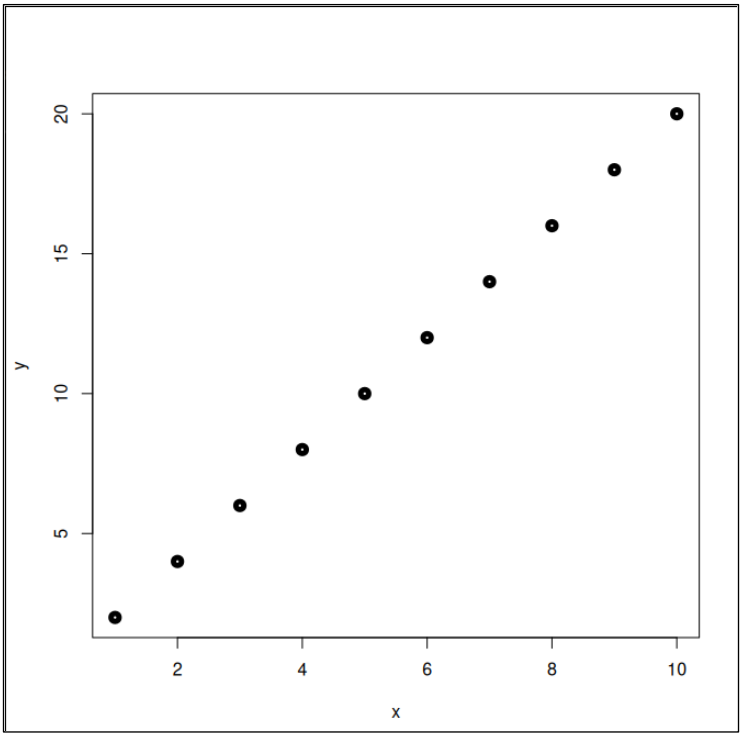


Figure 2.4.2 Example plot of x and y with correlation +1

Figure 2.4.3 Example plot of x and y with correlation 0.9728

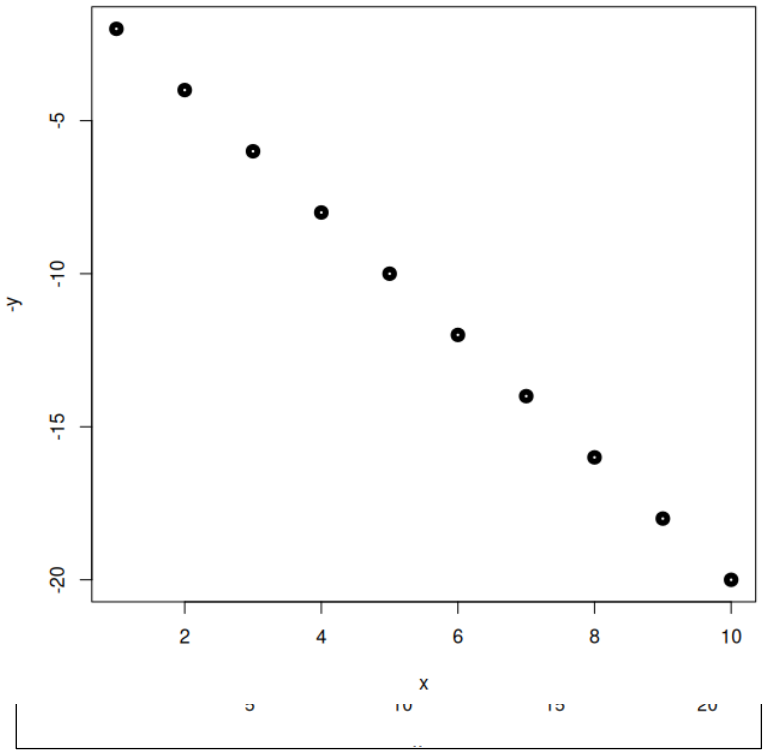


Figure 2.4.4 Example plot of x and y with correlation -1

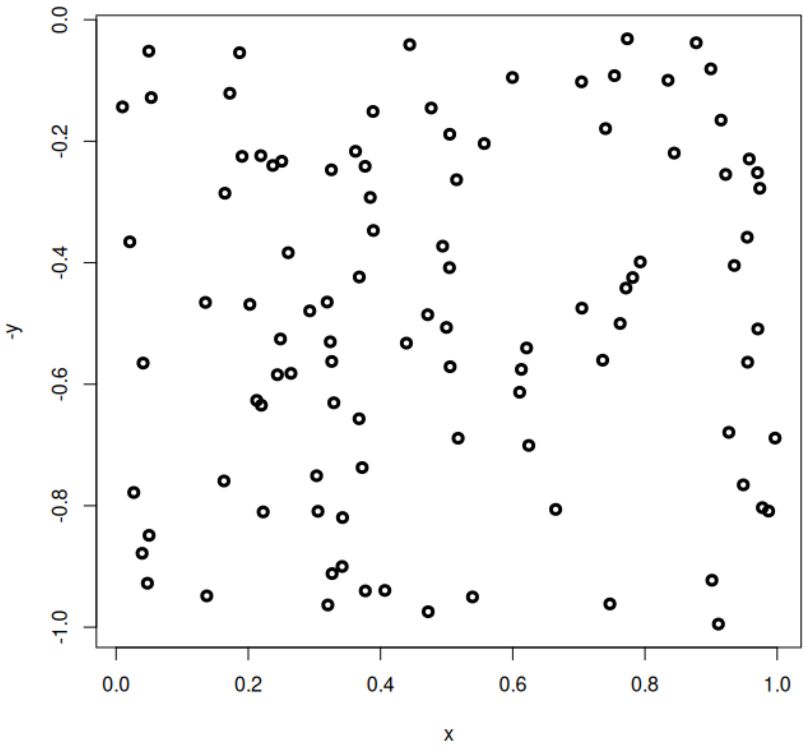


Figure 2.4.5 Example plot of x and y with correlation 0

The formula for Pearson correlation coefficient r is:

$$r = \frac{\Sigma(x_i - \underline{x})\,(y_i - \underline{y})}{\sqrt{\Sigma(x_i - \underline{x})^{\,2}\Sigma(y_i - \underline{y})^2}}$$

Where:

$x_i \; and \; y_i$ are the individual sample points.

$\underline{x} \;\; and \; \underline{y}$ are the mean of the x and y variables, respectively.

Example

Let's consider an example where we want to examine the correlation between hours studied and exam scores among a group of students.

Data: The sample data of exam score and study hours of seven students are given below.

Hours studied: 2, 3, 4, 5, 6, 7, 8

Exam scores: 50, 55, 60, 65, 70, 75, 80

The plot is given below.

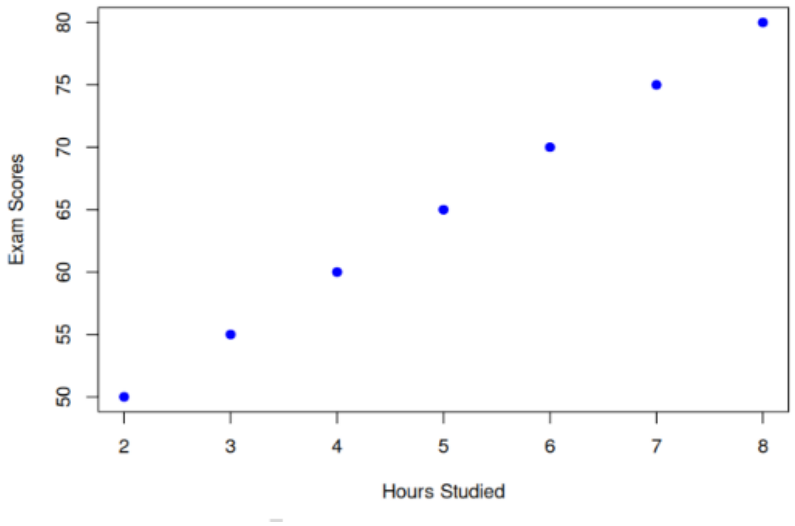


Figure 2.4.6 Scatter plot of Hours Studied vs Exam Scores

Pearson correlation coefficient between hours Studied and exam score is 1.

In this example, we expect a positive correlation because, generally, more hours studied should result in higher exam scores.

Problem

You are given a dataset with the following two variables: height (in inches) and weight (in pounds) of a group of people. Calculate the Pearson correlation coefficient and create a scatter plot to visualize the relationship.

Height: 60, 62, 64, 66, 68, 70, 72

Weight: 115, 120, 130, 140, 150, 160, 170

Solution:

Pearson Correlation Coefficient: 0.99734

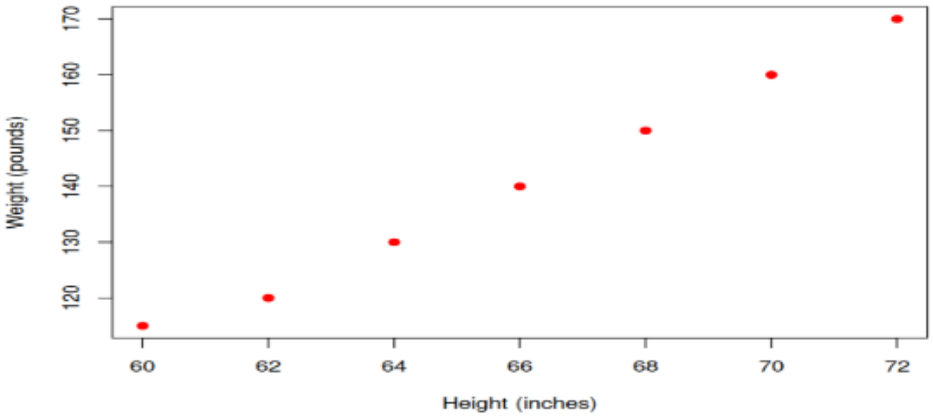


Figure 2.4.7 Scatter plot of Height vs Weight

When the data points are more scattered, then the correlation coefficient will more decreased. Note that correlation of x and y is same as that of y and x.

The following is the first 6 rows of 150 rows of measurements of petal and sepal of various iris flowers.

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
1	5.1	3.5	1.4	0.2
2	4.9	3.0	1.4	0.2
3	4.7	3.2	1.3	0.2
4	4.6	3.1	1.5	0.2
5	5.0	3.6	1.4	0.2
6	5.4	3.9	1.7	0.4

Figure 2.4.8 Measurements of petals and sepals

We can calculate the correlation coefficients of the columns.

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
Sepal.Length	1.0000000	-0.1175698	0.8717538	0.8179411
Sepal.Width	-0.1175698	1.0000000	-0.4284401	-0.3661259
Petal.Length	0.8717538	-0.4284401	1.0000000	0.9628654
Petal.Width	0.8179411	-0.3661259	0.9628654	1.0000000

Figure 2.4.9 calculation of coefficients of the columns

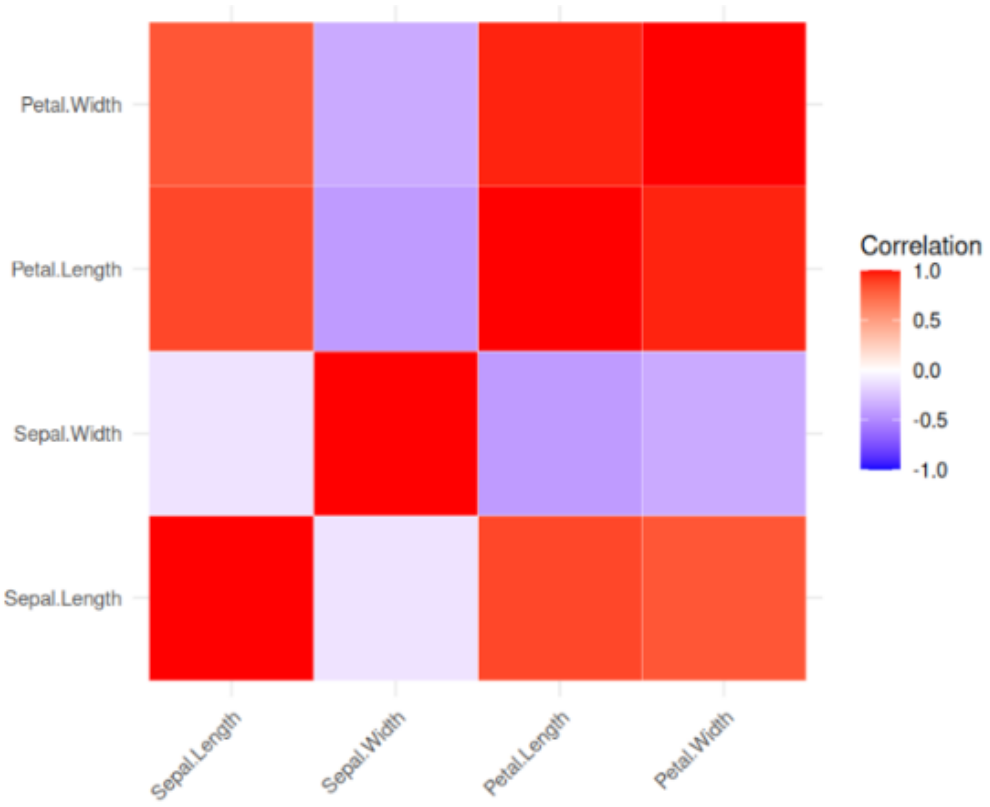


Figure 2.4.10 Correlation Matrix Heatmap of iris dataset

The above matrix can be plotted as heat map.

Correlation is a fundamental concept in statistics that helps us understand the relationship between two variables. The Pearson correlation coefficient provides a measure of the strength and direction of this relationship. Visualizing these relationships with scatter plots and heatmaps can provide valuable insights.

2.4.7 Graphs for Summarizing Data

2.4.7.1 Scatter Plots

A scatter plot is a simple yet powerful way to visualize the relationship between two quantitative variables. Each point on the plot represents an observation from the dataset.

Example:

Consider the iris data set we used in the previous example. Let us plot Petal.width versus Petal.Length

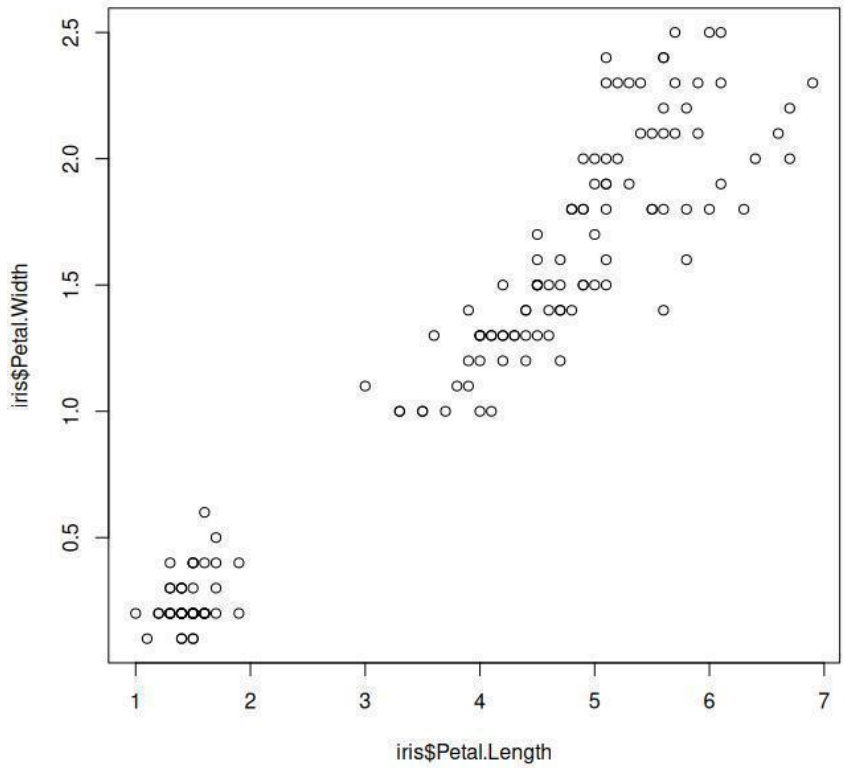


Figure 2.4.11 Plot of Petal.Width vs Petal.Length

This scatter plot gives an idea about how Petal.Length and Petal.Width vary together. When dealing with more than two variables, pair plots (or scatterplot matrices) are useful. They show scatter plots for all pairs of variables in a dataset.

Next, we plot all the four variables together, that is, the scatterplot matrix.

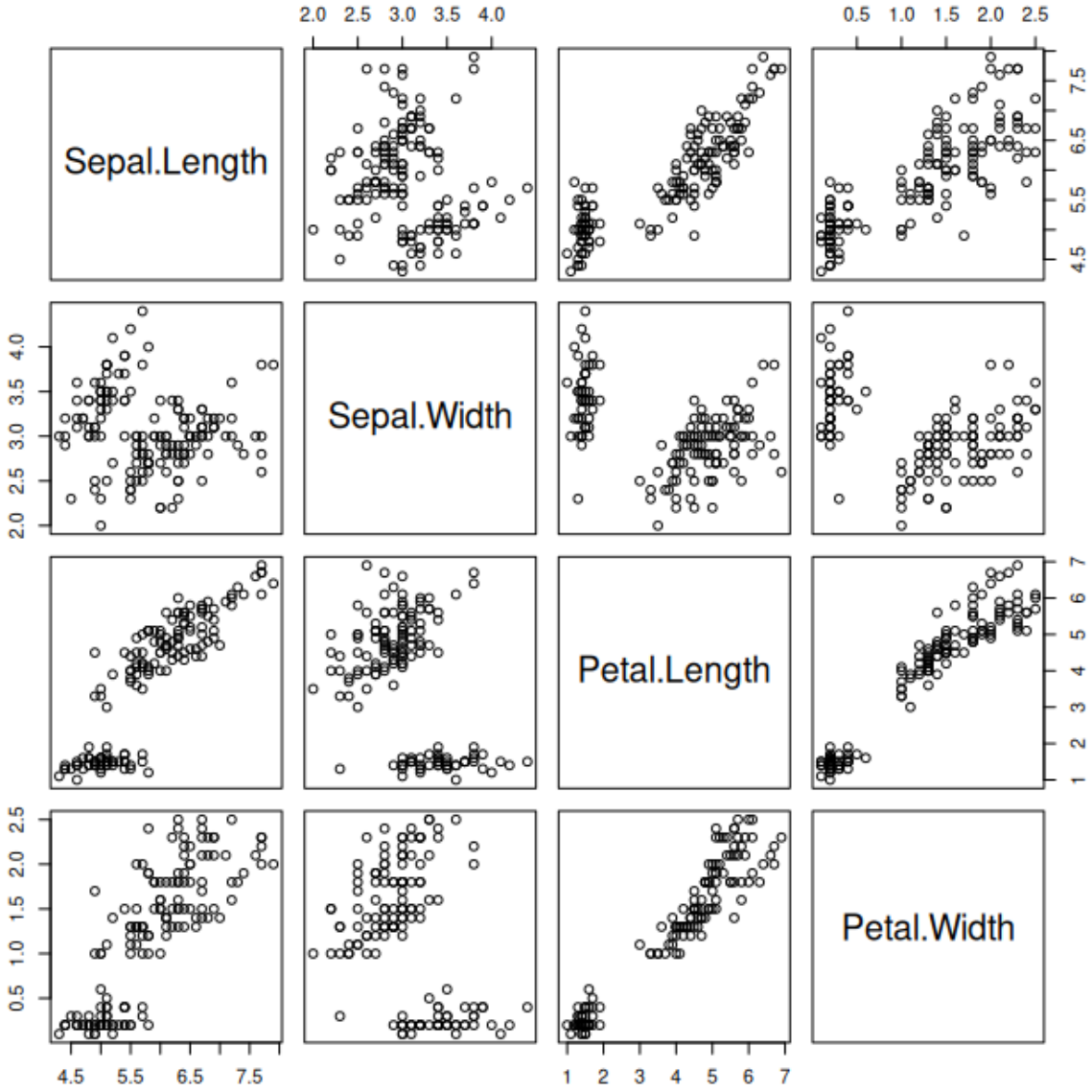


Figure2.4.12 Scatterplot Matrix of iris dataset

Summarizing data from more than one variable helps us understand the relationships and interactions between them. Scatter plots, pair plots, and heatmaps are valuable tools for visualizing these relationships. Correlation analysis provides a numeric measure of the strength and direction of these relationships. With practice and the use of R, you can effectively analyse and interpret multi-variable data.

Recap

- **Counting:** Tracks frequency of events or data points in a stream; crucial for real-time analysis.

➢ **Coefficient of Variation (CV)**

- Standardized measure of dispersion, expressed as a percentage.
- Calculated as the ratio of the standard deviation to the mean.
- Useful for comparing variability across datasets with different units or scales.
- Helps in understanding relative risk in finance and normalizing data for comparison.

➢ **Sampling Distribution of the Sample Mean**

- Describes how the sample mean varies from sample to sample.
- Mean of the sampling distribution equals the population mean.
- Standard error is the population standard deviation divided by the square root of the sample size.
- Approaches a normal distribution as sample size increases, according to the Central Limit Theorem.

➢ **Sampling Distribution of the Sample Proportion**

- Describes how the sample proportion varies from sample to sample.
- Mean of the sampling distribution equals the population proportion.
- Standard deviation of sampling distribution (standard error) is calculated by

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

- Approximates a normal distribution for large sample sizes.

➢ **Summarizing Data from More Than One Variable**

- **Graphs**: Used to visualize relationships between variables.

○ **Scatter Plots**: Show the relationship between two quantitative variables.

○ **Pair Plots**: Display scatter plots for all pairs of variables in a dataset.

- **Correlation**: Measures the strength and direction of the relationship between two variables.

○ **Pearson Correlation Coefficient**: Ranges from -1 to 1, indicating the degree of linear relationship.

- Provides insight into how variables move together and helps in making predictions.

Probability Theory and Distributions

UNIT 1
Probability Theory

Learning outcome

After completing this unit, the learner will be able to:

- Understand fundamentals of probability
- Familiarise probability definitions
- Familiarise probability rules
- Problem-Solving and Critical Thinking.

Pre-requisites

To calculate probabilities effectively, a basic understanding of set theory, permutations, and combinations is crucial, particularly in the classical approach, which assumes all outcomes are equally likely. The empirical approach relies on familiarity with data collection and interpreting frequencies. Rooted in logical reasoning, the axiomatic approach requires a grasp of mathematical definitions and rules.

Probability is the study of uncertainty and how likely events are to occur. It plays a crucial role in decision-making, risk assessment, and data analysis. The foundation of probability lies in understanding the sample space and events. For example, when flipping a coin, the sample space is {Heads, Tails}, and getting Heads is an event. Events can be simple, like rolling a 4 on a die, or compound, like rolling an even number. They can also be independent another, like flipping two different coins or dependent, like drawing cards from a deck without replacement. Rolling a 3 and a 5 on a single die is mutually exclusive. Probability also extends to random variables and their distributions, which help model real-life uncertainties, such as predicting weather conditions, stock market fluctuations, or medical diagnoses. Understanding these concepts allows us to quantify uncertainty and make informed decisions in various fields, from gaming and finance to science and technology.

Key Concepts

Probability, Sample space, Random variable, Mutually Exclusive Events

Discussion

Probability is a branch of mathematics that studies uncertainty and helps us understand the likelihood of different events happening. It provides a framework for predicting outcomes in situations where the result is not guaranteed, such as flipping a coin or rolling a die. By assigning numerical values to these chances, probability allows us to analyse and make decisions based on random events, whether they are simple or complex.

Consider the case of tossing a coin. Nobody knows whether the result is head or tail. But it is certain that a head or tail will occur. In a similar way, if a die is thrown, we may get any of the faces 1, 2, 3, 4, 5 and 6. But nobody knows which one will actually occur. Experiment of this type where the outcome cannot be predicted are called random experiment. The theory of probability analyses the result obtained by such experiments.

In this section we will define and explain the various terms which are used in the definition of probability.

3.1.1 The Concept of Probability

In statistics, probability serves as a fundamental concept, offering a quantitative measure of the uncertainty linked to events in a random experiment. To express this uncertainty, we assign a probability value ranging from 0 to 1. A value of 1 represents complete certainty that the event will occur, while a value of 0 indicates certainty that it will not. For example, if the probability is 1/4, we interpret it as a 25% chance of occurrence and a 75% chance of non-occurrence.

This numerical assignment enables us to quantify and convey our expectations about the likelihood of different outcomes.

Expressing probability numerically and understanding its implications have practical applications across various fields, such as risk assessment and decision-making. From evaluating the chances of success in a business venture to predicting outcomes in games of chance or making informed choices under uncertain conditions, probability and its numerical representation provide a crucial tool for quantifying and managing uncertainty in a wide range of scenarios.

Definitions

Trail and Event

Consider an experiment of tossing a coin. Here tossing a coin is a trail and getting a head or tail is an event.

Outcome

An outcome is a single result of an experiment.

For example, rolling a die and getting 4. That 4 is one outcome.

Event

An event is a collection of one or more outcomes.
For example, rolling a die and getting an odd number {1, 3, 5} is an event.

Single Event

An outcome is a single event.
For example, rolling a die and getting a 4 is a single event because it focuses on one specific outcome.

Compound Event

A compound event is an event that consists of two or more outcomes.
For example, rolling a number greater than 3 {4, 5, or 6}.

Exhaustive Event

A complete set of events that encompass all possible outcomes of a particular experiment.
For example, when flipping a coin, the events "heads" and "tails" are exhaustive because one of these two outcomes must happen in every flip.

Mutually Exclusive Event

Two events are called mutually exclusive when the events cannot occur at the same time. If one event happens, the other cannot happen.

For example, in throwing a die all 6 faces numbered 1 to 6 are mutually exclusive since if anyone of these faces comes, the possibility of others, in the same trial, is ruled out.

Equally Likely

Two or more events are said to be equally likely if the probability of each event occurring is the same.

For example, in tossing a fair coin, getting head or tail are equally likely events as both have the same chance of occurring.

Independent Events

Two events are said to be independent when the actual happening of one does not influence in any way happening of the other.

For example, outcome of tossing a coin (Head or Tail) does not affect the outcome of rolling a die (1 through 6) and vice versa.

3.1.2 Probability

Probability of happening an event $E = \frac{\text{Number of Favourable cases}}{\text{Total Number of Cases}}$

Probability of happening of an event is p and probability of not happening of an event is q and $p + q = 1$.

Probability values are always assigned on a scale from 0 to 1. A probability near zero indicates an event is unlikely to occur; a probability near 1 indicates an event is almost certain to occur. Other probabilities between 0 and 1 represent degrees of likelihood that an event will occur.

For example, if the weather report states a "near-zero probability of rain", it suggests that there is almost no chance of rain. However, a reported probability of "0.90" indicates that rain is highly likely. A probability of "0.50" means that rain is equally likely to happen or not happen.

Illustration 3.1.1

Find the probability of getting a) 4 b) an odd number c) an even number with an ordinary dice.

Solution

The total number of possible outcomes when rolling a die is 6.

a) When rolling a die, there is only one possible outcome to get a 4.
∴ The probability of rolling a 4 = 1/6

b) Number of ways of rolling {1, 3, 5} is 3

∴ Probability of getting an odd number $= \frac{3}{6} = \frac{1}{2}$

c)Number of ways of rolling {2, 4, 6} is 3

∴ Probability of getting an even number $= \frac{3}{6} = \frac{1}{2}$

Illustration 3.1.2

Find the probability of getting a sum of 8 with two dice.

Solution

The total number of possible ways of throwing 2 dice is $6 \times 6 = 36$.

Number of ways of getting 8 as sum $=(2 + 6), (3 + 5), (4 + 4), (5 + 3), (6 + 2) = 5$

∴ Probability of getting $8 = 5/36$

Illustration 3.1.3

A box contains 6 red and 7 black balls. Find the probability of drawing a red ball.

Solution

The total number of possible ways of getting a ball = 6+7 =13

Number of ways of getting 1 red ball = 6

∴ Probability of drawing a red ball = 6/13

Illustration 3.1.4

A bag contains 7 white, 6 red and 5 black balls. 2 balls are drawn at random. Find the probability that they both will be white.

Solution

The total number of balls $= 7 + 6 + 5 = 18$

From these 18 balls 2 balls can be drawn in $18C_2$ ways $= \frac{18\times17}{1\times2} = 153$

∴ The total number of ways of drawing 2 balls = 153.

2 white balls can be drawn from 7 white balls in $7C_2$ ways = 21 ways.

Probability of drawing 2 white balls $= \frac{21}{153} = \frac{7}{51}$.

Illustration 3.1.5

From a group of 3 Indians, 4 Pakistanis and 5 Americans, a subcommittee of 4 people is selected by lots. Find the probability that the subcommittee will consists of 2 Indians and 2 Pakistanis.

Solution

The total number of people $= 3 + 4 + 5 = 12.$

4 people can be chosen from 12 people in $12C_4$ ways.

i.e., $\frac{12!}{4!\times8!} = \frac{12\times11\times10\times9}{1\times2\times3\times4} = 495$ ways.

2 Indians can be chosen from 3 Indians in $3C_2$ ways.

2 Pakistanis can be chosen from 4 Pakistanis in $4C_2$ ways.

Number of ways of choosing 2 Indians and 2 Pakistanis in $3C_2 \times 4C_2$ ways.

∴ probability that the sub committee will consists of 2 Indians and 2 Pakistanis

$$= \frac{3C_2\times4C_2}{495} = \frac{6}{165}$$

3.1.3 Sample Space

A sample space can be defined as the list of all possible outcomes of a random experiment. It is denoted by S

For example, consider the toss of 2 coins. The sample space is

$$S = \{(HH), (HT), (TH), (TT)\}$$

Sample space is of two types. Finite sample space and Infinite sample space. In tossing a coin the sample space $S = \{H, T\}$ is a finite sample space.

3.1.4 Random Variable

Consider 2 tosses of a coin. Let the sample space be $S = \{(HH), (HT), (TH), (TT)\}$

Let X is the number of heads in each throw. The X takes the values $\{0, 1, 2\}$.

∴ X is a variable defined over the sample space of a random experiment called random variable. We denote random variables by capital letters and any particular value of random variables by small letters.

i, e. Random variable is a variable defined over the sample space of a random experiment.

There are two types of random variables.

1) Discrete random variable: Random variable whose possible values form a finite or countably infinite set of values.
   For example, tossing a coin and finding the number heads

2) Continuous random variable: Random variable which can assume any value within the interval.
   For example, age, height, weight are continuous random variables.

Recap

- Outcome: A single result of an experiment.
- Event: A collection of one or more outcomes.
- Mutually Exclusive Events: Two events are called Mutually exclusive when the occurrence of one affects the occurrence of the other.
- Independent Events: Two events are said to be independent when the actual happening of one does not influence in any way happening of the other.
- Probability $= \frac{\text{Number of Favourable cases}}{\text{Total Number of Cases}}$.
- Random Experiments: Experiments that can be performed many times under the same conditions and their outcome cannot be predicted with complete certainty
- Sample Space: The list of all possible outcomes of a random experiment. It is denoted by S
- Random Variable: a variable defined over the sample space of a random experiment

Objective Type Questions

1. What is an outcome in probability?
2. What is an event in probability?
3. Random variable which can assume any value within the interval.
4. When are two events called mutually exclusive?
5. The occurrence of one event does not influence the other then the events is called.
6. Name the process with unpredictable outcomes performed under the same conditions.
7. What is the range of probability values?
8. What is random variable?
9. What is the main characteristic of random experiments?
10. What is sample space of an experiment?

Answer to Objective Type Questions

1. A single result of an experiment.
2. A collection of one or more outcomes.
3. Continuous random variable.
4. When they cannot happen at the same time.
5. Independent events.
6. Random experiment.
7. 0 to 1
8. A variable defined over the sample space of a random experiment.
9. Their outcomes cannot be predicted with complete certainty.
10. The set of all possible outcomes of a random experiment.

Objective Questions

1. A fair six-sided die is rolled. List the sample space and find the probability of rolling:
   a) A 4 b) An even number c) A number greater than 4.
2. A bag contains 5 red, 3 blue, and 2 green balls. One ball is drawn at random. Find the probability that it is a) Red b) Not blue c) Either red or green
3. A coin is flipped, and a die is rolled. Find the probability that the coin lands on heads and the die shows a 6.
4. A box contains 3 red and 5 black balls. Two balls are drawn without replacement. Find the probability that both balls are red.
5. A card is drawn from a standard deck of 52 cards. Find the probability that it is either a King or an Ace.

Answers

1. $\{1,2,3,4,5,6\}$, a) $\frac{1}{6}$, b) ½ c) $\frac{1}{3}$
2. $\frac{1}{2}, \frac{7}{10}, \frac{7}{10}$
3. $\frac{1}{12}$
4. $3/28$
5. $\frac{2}{13}$

Suggested reading

1. Huffman, K., & Kunze, R. (2005). *Linear algebra* (2nd ed.). Pearson Education.
2. Apostol, T. M. (1991). *Calculus* (2nd ed.). Wiley. ISBN 960-07-0067-2.
3. Spivak, M. (2008). *Calculus*. Publish or Perish. ISBN 978-0914098911.
4. Finney, R. L., Weir, M. D., & Giordano, F. R. (2009). *Thomas's calculus* (12th ed.). Pearson.
5. Bendat, J. S., & Piersol, A. G. (2010). *Random data: Analysis and measurement procedures* (4th ed.). Wiley.

References

1. Kolman, B., Busby, R. C., Ross, S. T., & Rehman, M. (2017). *Discrete mathematical structures for computer science* (6th ed.). Pearson Education.
2. Kreyszig, E. (2011). *Advanced engineering mathematics* (9th ed.). Wiley India.
3. Ross, S. (2019). *A first course in probability* (9th ed.). Pearson Education.
4. Durrett, R. (2019). *Probability: Theory and examples* (5th ed.). Cambridge University Press.
5. Durrett, R. (1999). *Elementary probability for applications*. W. H. Freeman and Company

UNIT 2

Rules of Probability: Addition and Multiplication

Learning outcome

After completing this unit, learners will be able to:

- Familiarise the addition laws of probability
- Discuss multiplication laws of probability
- Narrate how to calculate the likelihood of single and combined events, including mutually exclusive and independent events
- Familiarise how to solve real-world problems

Pre-requisites

Imagine you are at a fun carnival, playing games where you receive either a red or blue ticket as a prize. You wonder what your chances are if you play two games and get one red ticket from one game and one blue ticket from the other. At first, you can calculate the chances of winning in each game separately. But what if you want to know the combined chance of winning from both games? This is where the addition rule comes in. It helps you add the chances of different events, like winning from either game. If the events cannot happen together (like getting a red ticket in one game and a blue ticket in another), the addition rule helps you figure out the total probability. If you are interested in winning both prizes, you use the multiplication rule. This rule tells you the chance of two events happening together, such as drawing a red ticket from one game and a blue ticket from another. By using these two rules, you can easily determine your chances of winning at the carnival games.

Keywords:

Addition law, Multiplication law

Discussion

The addition and multiplication laws of probability are key concepts that help us calculate the likelihood of different events. The addition rule is used to find the probability of one event or another happening, while the multiplication rule helps determine the probability of two events

occurring together. These rules are crucial for solving more complex probability problems, as they allow us to combine simple probabilities to predict the outcome of multiple events.

3.2.1 Addition law of probability

Let S be the sample space of a given experiment. Let A and B be two events of S. A∪B represents the event where either event A, event B, or both occur during the experiment. A∩B represents the event where both event A and event B happen simultaneously.

Then, $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

This rule can be extended to three or more events, for example:

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)$$

If two events A and B are mutually exclusive, then

$$P(A \cup B) = P(A) + P(B)$$

3.2.2 Multiplication Law of probability

If A and B are independent events, then

$$P(A \cap B) = P(A) \times P(B)$$

i.e. The probability of independent events A and B occurring is the product of the probabilities of the events occurring separately.

For example, If the probability of rolling a 3 on a fair die is $\frac{1}{6}$ and the probability of flipping a head on a fair coin is $\frac{1}{2}$, then:

P (rolling a 3 and flipping a head) $= \frac{1}{6} \times \frac{1}{2} = \frac{1}{12}$

Illustration 3.2.1

A person is known to hit the target in 3 out of 4 shots, where as another person is known to hit the target in 2 out of 3 shots. Find the probability that the target is hit when both persons try.

Solution

Probability of the first person hit the target = $P(A) = \frac{3}{4}$

Probability of the second person hit the target = $P(B) = \frac{2}{3}$

The two events are not mutually exclusive since both persons hit the same target.

∴ Required probability P (A or B) = $P(A) + P(B) - P(A \cap B)$

As it is an independent event, $P(A \cap B) = P(A) \times P(B)$

$$= \left(\frac{3}{4} + \frac{2}{3}\right) - \left(\frac{3}{4} \times \frac{2}{3}\right)$$

$$= \frac{17}{12} - \frac{6}{12}$$

$$= \frac{11}{12}$$

Illustration 3.2.2

A bag contains 20 balls,3 are coloured red, 6 are coloured green, 4 are coloured blue, 2 are coloured white and 5 are coloured yellow. One ball is selected at random. Find the probabilities of the following events. (a) the ball is either red or green (b) the ball is not blue (c) the ball is either red or white or blue.

Solution

P (getting red ball) = $\frac{3}{20}$ P (getting green ball)= $\frac{6}{20}$ P (getting blue ball)= $\frac{4}{20}$

P (getting white ball)= $\frac{2}{20}$

a) P (the ball is either red or green) = $\frac{3}{20} + \frac{6}{20} = \frac{9}{20}$

b) P (the ball is blue)= $\frac{4}{20}$, P (the ball is not blue) = $1 - \frac{4}{20} = \frac{16}{20} = \frac{4}{5}$

c) P (the ball is either red or white or blue)= $\frac{3}{20} + \frac{2}{20} + \frac{4}{20} = \frac{9}{20}$.

Illustration 3.2.3

From a pack of cards, a single card is drawn. What is the probability that it is either spade or king?

Solution

$P(A) = P(\text{Spade card}) = 13/52$

$P(B) = P(\text{King card}) = 4/52$

$P(\text{either spade or king}) = P(A) + P(B) - P(A \cap B)$

$$= \frac{13}{52} + \frac{4}{52} - \frac{13}{52} \times \frac{4}{52}$$

$$= \frac{13}{52} + \frac{4}{52} - \frac{52}{52 \times 52}$$

$$= \frac{16}{52}$$

$$= \frac{4}{13}$$

Illustration 3.2.4

The probability that machine A will be performing a usual function in 5 years' time is 1/4 while the probability that the machine B will still be operating usefully at the end of the same period is 1/3. Find the probability that both machines will be performing a usual function.

Solution

P (machine A operating usually)$= \frac{1}{4}$

P (machine B operating usually)$= \frac{1}{3}$

P (both machines will be performing usual function)$= P(A) \times P(B) = \frac{1}{4} \times \frac{1}{3} = \frac{1}{12}$

Illustration 3.2.5

A bag contains 8 white and 10 black balls. Two balls are drawn in succession. What is the probability that first is white and the second is black?

Solution

Total number of balls = 8+10 = 18

P (drawing a white ball) = 8/18

P (drawing a black ball) = 10/18

P (drawing first white and second black ball)$= \frac{8}{18} \times \frac{10}{18} = \frac{20}{81}$

Illustration 3.2.6

What is the chance of getting 2 sixes in twice rolls of a single dice?

Solution

P (a six in the first rolling) = 1/6

P (a six in the second rolling) = 1/6

As 2 rolls are independent, the probability of getting 2 sixes in 2 rolls $= \frac{1}{6} \times \frac{1}{6} = \frac{1}{36}$

Illustration 3.2.7

A bag contains 6 white balls and 9 black balls. Three drawings of one ball each are made, such that:
i) The balls are replaced before the next drawing, and
ii) The balls are not replaced before the next drawing.
Find the probability that all three drawn balls are black in each case.

Solution

i) P (drawing a black ball)$= \frac{9}{15}$

If the balls are replaced before the next drawing is made, the drawing of consecutive balls are independent and probability of each of them being black is 9/15.

So, the probability of all the three balls being black = $\frac{9}{15} \times \frac{9}{15} \times \frac{9}{15} = \frac{729}{3375}$

ii) If the balls are not replaced before the next drawing is made, the sample space for second and third drawing changes.

For the first draw, P (drawing a black ball) $= \frac{9}{15}$

After one black ball is drawn, there are now 8 black balls and 14 total balls remaining.

So, P (drawing another black ball) $= \frac{8}{14}$

After two black balls are drawn, there are now 7 black balls and 13 total balls remaining.

So, P (drawing another black ball)$= \frac{7}{13}$

P (drawing 3 black balls without replacement) = $\frac{9}{15} \times \frac{8}{14} \times \frac{7}{13} = \frac{504}{2730}$

Illustration 3.2.8

From a bag containing 4 white and 6 black balls, 2 balls are drawn at random.
a) If the balls are drawn one after the other without replacement, find the probability that
i) both the balls are white ii) both the balls are black iii) the first ball is white, and the second ball is black and iv) one ball is white and the other is black.
b) Find the probability in each of the casas if the balls are drawn one after the other with replacement

Solution

a) Without replacement

i) Both balls are white

P (first ball is white) = $\frac{4}{10}$

P (second ball is white) = $\frac{3}{9}$

Required probability = $\frac{4}{10} \times \frac{3}{9} = \frac{2}{15}$

ii) Both balls are black

P (first ball is black) $= \frac{6}{10}$

P (second ball is black) $= \frac{5}{9}$

Required probability $= \frac{6}{10} \times \frac{5}{9} = \frac{1}{3}$

iii) The first ball is white, and the second ball is black

P (first ball is white) $= \frac{4}{10}$

P (second ball is black) $= \frac{6}{9}$

Required probability $= \frac{4}{10} \times \frac{6}{9} = \frac{4}{15}$

iv) One ball is white, and the other is black (order does not matter)

P (first ball is white) $= \frac{4}{10}$, P (second ball is black) $= \frac{6}{9}$

P (first ball is white, and second ball is black) $= \frac{4}{10} \times \frac{6}{9} = \frac{24}{90}$

P (first ball is black) $= \frac{6}{10}$, P (second ball is white) $= \frac{4}{9}$

P (first ball is black, and second ball is white) $= \frac{6}{10} \times \frac{4}{9} = \frac{24}{90}$

Required probability $= \frac{24}{90} + \frac{24}{90} = \frac{8}{15}$

b) With replacement

i) Both balls are white

P (first ball is white) $= \frac{4}{10}$

P (second ball is white) $= \frac{4}{10}$

Required probability $= \frac{4}{10} \times \frac{4}{10} = \frac{4}{25}$

ii) Both balls are black

P (first ball is black) $= \frac{6}{10}$

P (second ball is black) $= \frac{6}{10}$

Required probability $= \frac{6}{10} \times \frac{6}{10} = \frac{9}{25}$

iii) The first ball is white, and the second ball is black

P (first ball is white) $= \frac{4}{10}$

P (second ball is black) $= \frac{6}{10}$

Required probability $= \frac{4}{10} \times \frac{6}{10} = \frac{6}{25}$

iv) One ball is white, and the other is black (order does not matter)

P (first ball is white) $= \frac{4}{10}$, P (second ball is black) $= \frac{6}{10}$

P (first ball is white and second ball is black) $= \frac{4}{10} \times \frac{6}{10} = \frac{24}{100}$

P (first ball is black) $= \frac{6}{10}$, P (second ball is white) $= \frac{4}{10}$

P (first ball is black and second ball is white) $= \frac{4}{10} \times \frac{6}{10} = \frac{24}{100}$

Required probability $= \frac{24}{100} + \frac{24}{100} = \frac{12}{25}$

Recap

- Addition Law of Probability: It is used to find the probability that at least one of two events occurs.
- Multiplication law of probability: It is used to calculate the probability of two or more events occurring together.

Objective Type Questions

1. If two events A and B are mutually exclusive, what is the probability of A or B occurring?
2. The probability of event A is 0.5 and the probability of event B is 0.3. What is the probability that either A or B occurs, assuming A and B are independent?
3. If the probability of an event occurring is 0.4, what is the probability of the event not occurring?
4. A card is drawn from a standard deck of 52 cards. What is the probability of drawing a red card?
5. The probability that a student passes a test is 0.75. What is the probability that the student fails the test?

Answers

1. P(A) + P(B)
2. 0.8
3. 0.6
4. 4.1/2
5. 5.0.25

Assignment Questions

1. In a box, there are 5 red balls, 3 green balls, and 2 blue balls. Two balls are drawn from the box at random without replacement. Find the probability that: a) The first ball drawn is red and the second ball drawn is green.
   b) At least one of the two balls drawn is red.
   Ans: 1/6, 7/9
2. In a class of 30 students, 18 students are taking mathematics, 12 students are taking science, and 8 students are taking both subjects. A student is selected at random. Find the probability that the student: a) is taking mathematics or science. b) is taking neither mathematics nor science.

   Ans: 11/15, 4/15

3. A deck of 52 cards contains 13 hearts, 13 diamonds, 13 spades, and 13 clubs. Two cards are drawn consecutively from the deck without replacement. Calculate the probability that: a) Both cards drawn are red. b) The first card is a spade, and the second card is a heart.
   Ans:25/51,13/102
4. A company has 80 employees, of which 40 are male and 40 are female. 10 male and 5 female employees are selected at random for a meeting. Find the probability that: a) Both selected employees are male. b) At least one of the selected employees is female.
   Ans:39/158, 119/158
5. A fair die is rolled twice. Find the probability that: a) The sum of the two rolls is 7 or greater. b) The first roll is a 4 and the second roll is an odd number.

   Ans: 7/12, 1/12

Suggested reading

1. Huffman, K., & Kunze, R. (2005). *Linear algebra* (2nd ed.). Pearson Education.
2. Apostol, T. M. (1991). *Calculus* (2nd ed.). Wiley. ISBN 960-07-0067-2.
3. Spivak, M. (2008). *Calculus*. Publish or Perish. ISBN 978-0914098911.
4. Finney, R. L., Weir, M. D., & Giordano, F. R. (2009). *Thomas's calculus* (12th ed.). Pearson.
5. Bendat, J. S., & Piersol, A. G. (2010). *Random data: Analysis and measurement procedures* (4th ed.). John Wiley & Sons, Inc.

References

1. Montgomery, D. C., & Runger, G. C. (2011). *Applied statistics and probability for engineers* (5th ed.). John Wiley & Sons, Inc.
2. Kolman, B., Busby, R. C., Ross, S. T., & Rehman, M. (2017). *Discrete mathematical structures for computer science* (6th ed.). Pearson Education.
3. Kreyszig, E. (2011). *Advanced engineering mathematics* (9th ed.). Wiley India.
4. Ross, S. (2019). *A first course in probability* (9th ed.). Pearson Education.
5. Durrett, R. (2019). *Probability: Theory and examples* (5th ed.). Cambridge University Press.

Unit 3

Conditional probability and independence

Learning outcome

After completing this unit, learners will be able to:

- Define and calculate conditional probability
- Develop the ability to interpret conditional probability.
- Apply conditional probability
- Discuss the concept of independence to solve problems in various domains.

Pre-requisites

Conditional Probability refers to the probability of an event occurring given that another event has already occurred. Conditional probability is important when events are not independent, and we want to understand how the occurrence of one event impacts the likelihood of another event. A solid understanding of basic probability concepts, such as probability rules, the Addition and Multiplication Laws of Probability is essential to have a strong grasp of conditional probability concepts. A grasp of basic algebra and fractions also plays a role in simplifying and interpreting conditional probability expressions.

Keywords:

Conditional probability, Independent Events

Discussion

Suppose a bag contains 6 balls, 3 red and 3 white. Two balls are chosen (without replacement) at random, one after the other. Consider the two events, R is the event that the first ball chosen is red, W the event that the second ball chosen is white.

We easily find $P(R) = \frac{3}{6} = \frac{1}{2}$. If the first ball chosen is red, then the bag subsequently contains 2 red balls and 3 white. In this case $P(W) = \frac{3}{5}$. However, if the first ball chosen is white then the bag subsequently contains 3 red balls and 2 white. In this case $P(W) = \frac{2}{5}$. i.e, the

probability that W occurs is clearly dependent upon whether the event R has occurred. The probability of W occurring is conditional on the occurrence or otherwise of R.
The conditional probability of an event B occurring given that event A has occurred is written P(B|A). In this example $P(W|R) = \frac{3}{5}$ and $P(W|R') = \frac{2}{5}$. The probability of W occurring is conditional on the occurrence or otherwise of R.

3.3.1 Conditional probability

The conditional probability P(B|A) is defined as

$$P(B|A) = \frac{\text{number of outcomes in } A \cap B}{\text{number of outcomes in } A} = \frac{P(A \cap B)}{P(A)}$$

Or

$$P(A \cap B) = P(B|A) \times P(A)$$

Or

$$P(AB) = P(B|A) \times P(A)$$

3.3.2 Independent Events

A set of events is said to be independent events if the occurrence of any one of them does not depend on the occurrence or non-occurrence of the other.

If two events A and B are independent, then by $P(A \cap B) = P(A) \times P(B)$

Then $P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{P(A) \times P(B)}{P(A)} = P(B)$

so, the probability of B is unaffected by knowledge that A occurred.

For example, toss 2 fair coins and let F = head on 1st toss, S = head on 2nd toss. These are independent events.

Properties of Independence

If A and B are independent, then A and B′ are independent.

If A and B are independent, then A′ and B′ are independent.

Let events A, B, C be mutually independent. Then A and B∩C are independent, and A and B∪C are independent

Illustration 3.3.1
A bag contains 3 red ad 4 white balls. Two draws are made without replacement. What is the probability that both the balls are red?

Solution

P (drawing a red ball in the first draw), $P(A) = \frac{3}{7}$

P (drawing a red ball in the second draw given that first ball drawn is red), $P(B|A) = \frac{2}{6}$

$$P(AB) = P(B|A) \times P(A)$$
$$= \frac{2}{6} \times \frac{3}{7} = \frac{1}{7}$$

Illustration 3.3.2

Find the probability of drawing a queen and a king from a pack of cards in two consecutive draws, the cards drawn not being replaced?

Solution

P (drawing a queen card), $P(A) = \frac{4}{52}$

P (drawing a king after a queen has been drawn), $P(B|A) = \frac{4}{51}$

$$P(AB) = P(B|A) \times P(A)$$
$$= \frac{4}{52} \times \frac{4}{51} = \frac{4}{663}$$

Illustration 3.3.3

An urn contains 10 balls, 4 are red and the remaining green. What is the probability of drawing a red ball after drawing a green ball without replacement?

Solution

P (drawing a green ball in the first draw), $P(A) = \frac{6}{10}$

P (drawing a green ball given that first ball drawn is red), $P(B|A) = \frac{4}{9}$

$$P(AB) = P(B|A) \times P(A)$$
$$= \frac{4}{9} \times \frac{6}{10} = \frac{16}{90}$$

Illustration 3.3.4

In a survey among few people, 60% read Hindi newspaper, 40% read English newspaper and 20% read both. If a person is chosen at random and if he already reads English newspaper find the probability that he also reads Hindi newspaper.

Solution:

Let there be 100 people in the survey, then

Number of people read Hindi newspaper $= n(A) = 60$

Number of people read English newspaper $= n(B) = 40$

Number of people read both $= n(A \cap B) = 20$

Probability of the person reading Hindi newspaper when he already reads English newspaper is given by $P(A|B) = \frac{n(A \cap B)}{n(B)} = \frac{20}{40} = \frac{1}{2}$

Illustration 3.3.5

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

Solution:

Let A: event of selecting a red ball in first draw

B: event of selecting a red ball in second draw

$P(A \cap B) = P$ (selecting both red balls) $= \frac{3}{10} \times \frac{2}{9} = \frac{1}{15}$

P(B) = P (red ball on 1st and red ball on 2nd) + P (black ball on 1st and red ball on 2nd)

$$= \frac{3}{10} \times \frac{2}{9} + \frac{7}{10} \times \frac{3}{9} = \frac{3}{10}$$

$\therefore P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{1}{15} \text{x} \frac{3}{10} = \frac{2}{9}$s

Illustration 3.3.6

Suppose we flipped a coin and rolled a die. What is the probability of getting a head on the coin and a 6 on the die?

Solution

Let A = getting a head and B = rolling a 6

We know that $P(A) = \frac{1}{2} =$ and $P(B) = \frac{1}{6}$

A and B are independent since the probability of rolling a 6 does not depend on the outcome of the coin toss. Using the multiplication rule, we get:

$$P(A \text{ and } B) = P(A) \times P(B) = \frac{1}{2} \times \frac{1}{6} = \frac{1}{12}$$

Illustration 3.3.7

A bag contains 5 red and 4 white marbles. A marble is drawn from the bag, its color recorded, and the marble is returned to the bag. A second marble is then drawn. What is the probability that the first marble is red, and the second marble is white?

Solution

Since the first marble is put back in the bag before the second marble is drawn these are independent events.

$$P(1^{st} \text{ red and } 2^{nd} \text{ white}) = P(1^{st} \text{ red}) \times P(2^{nd} \text{ white})$$
$$= \frac{5}{9} \times \frac{4}{9} = \frac{20}{81}$$

Illustration 3.3.8

An urn contains 10 white, 3 black balls while another urn contains 3 white, 5 black balls. Two balls are drawn from the first urn and put into the second urn. Then a ball is drawn from the latter. What is the probability that it is a white ball?

Solution

Two balls drawn from the 1^{st} urn may be

i) both white (event A_1)
ii) both black (event A_2)
iii) 1 white, 1 black (event A_3)

$$\therefore P(A_1) = \frac{10C_2}{13C_2} = \frac{15}{26}$$

$$P(A_2) = \frac{3C_2}{13C_2} = \frac{1}{26}$$

$$P(A_3) = \frac{10C_1 \times 3C_1}{13C_2} = \frac{10}{26}$$

After the balls are transformed from 1^{st} urn to 2^{nd} urn, the 2^{nd} urn will contain
i) 5 white, 5 black ii) 3 white, 7 black iii) 4 white, 6 black
Let B be the event of drawing a white ball from the 2^{nd} urn.
Now $P(B|A_1) = \frac{5C_1}{10C_1} = \frac{5}{10}$

$$P(B|A_2) = \frac{3C_1}{10C_1} = \frac{3}{10}$$

$$P(B|A_3) = \frac{4C_1}{10C_1} = \frac{4}{10}$$

$$\therefore P(B) = \sum_{i=1}^{3} P(B|A_i)\, P(A_i)$$

$$= \frac{5}{10} \times \frac{15}{26} + \frac{3}{10} \times \frac{1}{26} + \frac{4}{10} \times \frac{10}{26} = \frac{59}{130}$$

Recap

- The conditional probability P(B|A) is defined as

$$P(B|A) = \frac{\text{number of outcomes in } A \cap B}{\text{number of outcomes in A}} = \frac{P(A \cap B)}{P(A)}$$

- A set of events is said to be independent events if the occurrence of any one of them does not depend on the occurrence or non-occurrence of the other

Objective Type Questions

1. Probability of an event occurring given that another event has already occurred is called.
2. If two events A and B are independent, what is the probability of A and B occurring together?
3. If $P(A|B) = 0.6$ and $P(B) = 0.5$, what is $P(A \cap B)$?
4. If the probability of event A is 0.4 and the probability of event B is 0.5, and A and B are independent, what is $P(A \cap B)$?
5. If the probability of event A occurring is 0.3 and the probability of event B occurring is 0.4, and events A and B are mutually exclusive, what is $P(A \cup B)$?
6. If $P(A|B) = 0.5$ and $P(B) = 0.4$, what is $P(A \cap B)$?
7. Two events A and B are independent. If $P(A) = 0.2$ and $P(B) = 0.3$, what is the probability that both events occur?
8. If $P(A \cap B) = 0.4$ and $P(A) = 0.6$, what is $P(B|A)$?
9. If the probability of event A is 0.9 and the probability of event B is 0.2, and A and B are independent, what is $P(A \cap B)$?
10. If $P(A|B) = 0.4$ and $P(B) = 0.8$, what is $P(A \cap B)$?

Anseres

1. Conditional Probability
2. $P(A \cap B) = P(A) \times P(B)$
3. 0.3
4. 0.2
5. 0.7
6. 0.2
7. 0.06
8. 0.67
9. 0.18
10. 0.32

Assignment

1. In a deck of 52 cards, what is the probability that a second card drawn is a heart given that the first card drawn was also a heart? **Ans: 4/17**
2. A fair die is rolled twice. What is the probability that the second roll is greater than 4 given that the first roll was a 3? **Ans: 1//3**
3. A class has 20 students, 12 are boys and 8 are girls. Two students are selected at random. What is the probability that the second student is a girl given that the first student selected was a boy?

   Ans: 8//19
4. A factory has 3 machines: A, B, and C. They produce 40%, 35%, and 25% of the total products, respectively. The probability that a product is defective from A, B, and C is 2%, 3%, and 4%, respectively. If a product is defective, what is the probability that it came from machine B?

   Ans: 0.3684
5. A coin is tossed. If heads appear, a die is rolled. If tails appear, two dice are rolled. What is the probability that the sum of the rolled numbers is 7 given that tails appeared?

   Ans:1/6
6. A student has a 70% chance of passing Math and an 80% chance of passing Science. If the probability of passing both is 60%, what is the probability that the student passes Science given that they passed Math? **Ans:0.8571**

Suggested reading

1. Huffman, K., & Kunze, R. (2005). *Linear algebra* (2nd ed.). Pearson Education.
2. Apostol, T. M. (1991). *Calculus* (2nd ed.). Wiley. ISBN 960-07-0067-2.
3. Spivak, M. (2008). *Calculus*. Publish or perish. ISBN 978-0914098911.
4. Finney, R. L., Weir, M. D., & Giordano, F. R. (2009). *Thomas's calculus* (12th ed.). Pearson.
5. Bendat, J. S., & Piersol, A. G. (2010). *Random data: Analysis and measurement procedures* (4th ed.). John Wiley & Sons, Inc.

References

1. Montgomery, D. C., & Runger, G. C. (2011). *Applied statistics and probability for engineers* (5th ed.). John Wiley & Sons, Inc.
2. Kolman, B., Busby, R. C., Ross, S. T., & Rehman, M. (2017). *Discrete mathematical structures for computer science* (6th ed.). Pearson Education.
3. Kreyszig, E. (2011). *Advanced engineering mathematics* (9th ed.). Wiley India
4. Ross, S. (2019). *A first course in probability* (9th ed.). Pearson Education.
5. Durrett, R. (2019). *Probability: Theory and examples* (5th ed.). Cambridge University Press.

Unit 4

Bayes' Theorem

Learning outcome

After completing this unit, learners will be able to:

- Apply Bayes' Theorem

Pre-requisites

Bayes' Theorem include a solid grasp of probability theory, specifically conditional probability, the probability of an event given that another has occurred, and the multiplication rule for independent events. Knowledge of basic set theory and operations such as union and intersection of events is also essential. Additionally, understanding prior probabilities (initial beliefs before new evidence) and likelihoods (the probability of observed data given a hypothesis) is necessary. Familiarity with probability distributions and total probability helps in applying Bayes' Theorem effectively, particularly when dealing with more complex scenarios.

Keywords

Bayes' Theorem

Discussion

3.4.1 Baye's theorem

Let $A_1, A_2, \ldots A_n$ be n exclusive and exhaustive events with $P(A_i) \neq 0$ for $i = 1,2, \ldots n$. Let B be an event such that $P(B) > 0$, Then the conditional probability of Ai given B is given by:

$$P(A_i|B) = \frac{P(A_i).\ P(B|A_i)}{\sum_{i=1}^{n} P(A_i).\ P(B|A_i)}$$

If A and B are two events, then $P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}$ where $p(B) \neq 0$

where P(A|B) is the probability of condition when event A is occurring while event B has already occurred.

Illustration.3.4.1

Out of 100 men, 5 are colourblind, and out of 1000 women, 25 are colourblind. A colourblind person is chosen at random. What is the probability that this person is male? (Assume that the number of males and females in the population is equal.)

Solution

Let M denote a person is Male. Let F denotes a person is Female. Let C denote a person is colour blind.

Given $P(M) = \frac{1}{2}$, $P(F) = \frac{1}{2}$ (Population is assumed to have an equal number of males and females)

Probability of being colourblind given the person is male, $P(C|M) = \frac{5}{100}$

Probability of being colourblind given the person is female, $P(C|F) = \frac{25}{1000}$

$$P(M|C) = \frac{P(C|M) \times P(M)}{P(C|M)P(M) + P(C|F)P(F)}$$

$$= \frac{\frac{5}{100} \times \frac{1}{2}}{\frac{5}{100} \times \frac{1}{2} + \frac{25}{1000} \times \frac{1}{2}}$$

$$= \frac{0.05}{0.05 + 0.025} = 2/3$$

Illustration.3.4.2

There are three identical boxes, labelled I, II, and III, each containing two coins. Box I hold two gold coins, Box II holds two silver coins, and Box III contains one gold and one silver coin. A person randomly selects a box and draws a coin, which turns out to be gold. What is the probability that the other coin in the selected box is also gold?

Solution

Let E_1, E_2 and E_3 be the events that boxes I, II and III are chosen, respectively.

Then $P(E_1) = P(E_2) = P(E_3) = \frac{1}{3}$

Also, Let A represent the event that "a gold coin is drawn"

Then

$P(A|E_1)$ = P (drawing a gold coin from box I) $= \frac{2}{2} = 1$

$P(A|E_2)$ = P (drawing a gold coin from box II) = 0

$P(A|E_3)$ = P (drawing a gold coin from box III) $= \frac{1}{2}$

Now, the probability of the other coin in the box is also gold= the probability of drawing a gold coin from box I.= $P(E_1|A)$

By Bayes theorem, we know that

$$P(E_1 \mid A) = \frac{P(E_1)P(A \mid E_1)}{P(E_1)P(A \mid E_1) + P(E_2)P(A \mid E_2) + P(E_3)P(A \mid E_3)}$$

$$= \frac{\frac{1}{3} \times 1}{\left(\frac{1}{3} \times 1\right) + \left(\frac{1}{3} \times 0\right) + \left(\frac{1}{3} \times \frac{1}{2}\right)} = \frac{2}{3}$$

Illustration.3.4.3

2% of the population have a certain blood disease in a serious form, 10% have it in a mild form, and 88% do not have it at all. A new blood test is developed, with the probability of testing positive being 9/10 for those with the serious form, 6/10 for those with the mild form, and 1/10 for those without the disease. I have just tested positive. What is the probability that I have the serious form of the disease?

Solution

Let A_1- disease in serious form,

A_2- disease in mild form,

A_3 - does not have disease,

B - test positive.

Then we are given that $P(A_1) = \frac{2}{100} = 0.02$, $P(A_2) = \frac{10}{100} = 0.1$, $P(A_3) = \frac{88}{100} = 0.88$

$P(B|A_1) = \frac{9}{10} = 0.9$, $P(B|A_2) = 0.6$, $P(B|A_3) = 0.1$

By Bayes' Theorem,

$$P(A_1|B) = \frac{P(B|A_1).P(A_1)}{P(B|A_1).P(A_1) + P(B|A_2).P(A_2) + P(B|A_3).P(A_3)}$$

$$= \frac{0.9 \times 0.02}{0.9 \times 0.02 + 0.6 \times 0.1 + 0.1 \times 0.88}$$

$$= \frac{0.018}{0.166} = 0.108$$

Illustration.3.4.4

A bag I contains 4 white and 6 black balls while another Bag II contains 4 white and 3 black balls. One ball is drawn at random from one of the bags, and it is found to be black. Find the probability that it was drawn from Bag I.

Solution:

Let E_1 be the event of choosing bag I, E_2 the event of choosing bag II, and A be the event of drawing a black ball.

$$P(E_1) = P(E_2) = \frac{1}{2}$$

$P(A|E_1) = P$ (drawing a black ball from bag I) $= \frac{6}{10}$

$P(A|E_2) = P$ (drawing a black ball from bag II) $= \frac{3}{7}$

By Bayes' theorem

$$P(E_1|A) = \frac{P(A|E_1).P(E_1)}{P(A|E_1).P(E_1) + P(A|E_2).P(E_2)}$$

$$= \frac{\frac{1}{2} \times \frac{6}{10}}{\frac{1}{2} \times \frac{6}{10} + \frac{1}{2} \times \frac{3}{7}}$$

$$= \frac{\frac{3}{10}}{\frac{3}{10} + \frac{3}{14}} = \frac{7}{12}$$

Illustration.3.4.5

An insurance company insured 2000 scooter drivers, 4000 car drivers, and 6000 truck drivers. The probability of an accident involving a scooter driver, car driver, and a truck is 0.01, 0.03, and 0.015, respectively. One of the insured persons meets with an accident. What is the probability that he is a scooter driver?

Let E_1, E_2, E_3, and A be the events defined as follows:

E_1 = person chosen is a scooter driver

E_2 = person chosen is a car driver

E_3 = person chosen is a truck driver and

A = person meets with an accident

Since there are 12000 people

$$P(E_1) = \frac{2000}{12000} = ⅙$$

$$P(E_2) = \frac{4000}{12000} = ⅓$$

$$P(E_3) = \frac{6000}{12000} = ½$$

Given that $P(A/E_1)$ = Probability that a person meets with an accident given that he is a scooter driver = 0.01

Similarly, you have $P(A/E_2) = 0.03$ and $P(A/E_3) = 0.015$

$$P(E_1|A) = \frac{P(A|E_1).P(E_1)}{P(A|E_1).P(E_1) + P(A|E_2).P(E_2) + P(A|E_3).P(E_3)}$$

$$= \frac{0.01 \times 1/6}{0.01 \times \frac{1}{6} + 0.03 \times \frac{1}{3} + 0.015 \times \frac{1}{2}}$$

$$= \frac{2}{23}$$

Objective-type questions

1. If P(A|B) = 0.7, P(B) = 0.5, and P (A ∩ B) = 0.35, what is P(B|A)?
2. Bayes' Theorem is useful for calculating which type of probability?
3. In Bayes' Theorem, the "prior probability" refers to:
4. If P(A) = 0.4, P(B|A) = 0.3, and P(B) = 0.5, what is P(A|B) according to Bayes' Theorem?
5. What is the "posterior probability" in the context of Bayes' Theorem?
6. If the probability of a disease is 0.02, the probability of a positive test given the disease is 0.95, and the probability of a positive test is 0.10, what is the probability of having the disease given a positive test result (P (Disease|Positive Test))?
7. In the context of Bayes' Theorem, if two events A and B are independent, then what is P(A|B)?
8. 8.What is “likelihood” in Bayes' Theorem?
9. If P(A) = 0.8, P(B|A) = 0.9, and P(B) = 0.4, what is P(A|B)?
10. 10.If P(A|B) = 0.6, P(B) = 0.2, and P (A ∩ B) = 0.75, what is P(B|A)?

Answer

1. 0.6
2. Conditional probability
3. The probability of observing the evidence before new information is considered.
4. 0.24
5. The probability of the event occurring after considering the evidence.
6. 0.19
7. P(A|B) = P(A)
8. Probability of observing the given data under a specific hypothesis
9. 1.8
10. 0.2

Assignment

1. An urn contains 7 white and 3 red balls. Two balls are drawn together, at random, from this urn. Compute the probability that neither of them is white. Also find the probability of getting one white and one red ball. Hence compute the expected number of white balls drawn.
   Ans: 1/15, 7/15, 1.4

2. An urn contains 10 white and 3 black balls. Another urn contains 3 white and 5 black balls. 2 balls are drawn at random from the first urn and placed in the second urn and then one ball is taken at random from the latter. What is the probability that it is a white ball?
   Ans: 59/130
3. A person is known to hit the target in 3 out of 4 shots whereas another person is known to hit the target in 2 out of 3 shots, Find the probability of the target being hit at all when both persons try.

   Ans: 11/12
4. An urn contains 5 balls. 2 balls are drawn and are found to be white. What is the probability of all the balls being white?

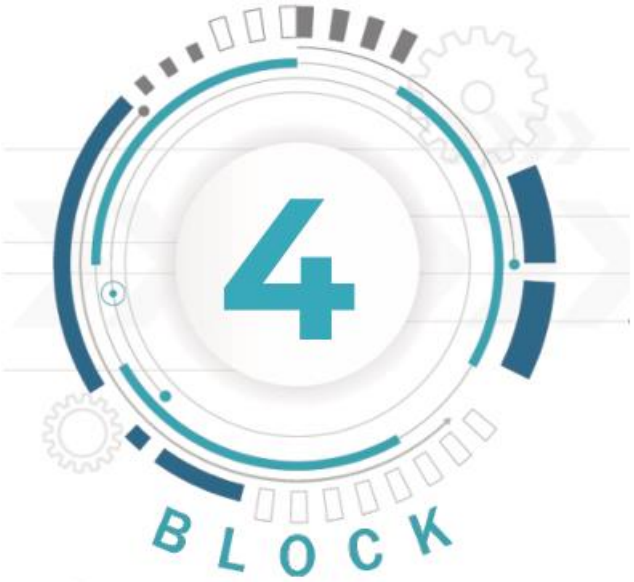
   Ans: 1/2
5. A man is known to speak the truth 2 out of 3 times. He throws a die and reports that the number obtained is a four. Find the probability that the number obtained is actually a four.
   Ans:2/7

Suggested reading

1. Montgomery, D. C., & Runger, G. C. (2011). *Applied statistics and probability for engineers* (5th ed.). John Wiley & Sons, Inc.
2. Kolman, B., Busby, R. C., Ross, S. T., & Rehman, M. (2017). *Discrete mathematical structures for computer science* (6th ed.). Pearson Education.
3. Kreyszig, E. (2011). *Advanced engineering mathematics* (9th ed.). Wiley India
4. Ross, S. (2019). *A first course in probability* (9th ed.). Pearson Education.
5. Durrett, R. (2019). *Probability: Theory and examples* (5th ed.). Cambridge University Press.

References

1. Huffman, K., & Kunze, R. (2005). *Linear algebra* (2nd ed.). Pearson Education.
2. Apostol, T. M. (1991). *Calculus* (2nd ed.). Wiley. ISBN 960-07-0067-2.
3. Spivak, M. (2008). *Calculus*. Publish or perish. ISBN 978-0914098911.
4. Finney, R. L., Weir, M. D., & Giordano, F. R. (2009). *Thomas's calculus* (12th ed.). Pearson.
5. Bendat, J. S., & Piersol, A. G. (2010). *Random data: Analysis and measurement procedures* (4th ed.). John Wiley & Sons, Inc.



Discrete and Continuous Probability Distributions

UNIT 1

Discrete Probability Mass Function, Binomial and Poisson distributions

Learning outcome

At the end of this unit, the student will be able to:

- understanding Probability Distribution
- Analyze and Solve Problems Using the Binomial Distribution
- Interpret and utilize the Poisson Distribution
- Differentiate and Choose Appropriate Discrete Distributions

Pre-requisites

In sports analytics, the Binomial Distribution helps predict the probability of a basketball player making a certain number of successful free throws in a game, based on their shooting accuracy. Familiarity with descriptive statistics is important for interpreting parameters like mean and variance, which are used in quality control to determine the likelihood of defective products in a batch. A knowledge of combinatorics is crucial for understanding how events are counted in binomial experiments, such as calculating the probability of getting exactly 3 heads in 5 coin tosses. Similarly, the Poisson Distribution is widely used in traffic flow analysis, where it models the number of cars passing through a toll booth in a given time period.

Additionally, a foundation in basic algebra and graphing functions helps in visualizing probability distributions. For instance, in hospital emergency room management, understanding the Poisson Distribution curve aids in predicting the expected number of patient arrivals per hour, ensuring proper resource allocation. By mastering these fundamental concepts, one can effectively apply probability distributions to real-world problems in business, healthcare, engineering, and beyond.

Key Concepts

Probability Mass function, Binomial, Poisson Distributions

Discussion

4.1.1 Discussion

Consider an experiment of throwing a coin twice. The outcomes {HH, TT, HT, TH} constitute the Sample space. Let X be the number of heads in each throw. Then X can take the values 0, 1, and 2.

Event	HH	HT	TH	TT
Value	2	1	1	0

X is a variable defined over the sample space of a random experiment called Random variable.

There are two types of random variables, discrete random variables and continuous random variables. A discrete random variable takes on a countable number of distinct values, such as whole numbers. Examples include the number of heads in a coin toss or the number of customers in a store.The probability function of a discrete random variable is probability mass function(PMF).

In the above example, the probability of getting 0 heads (TT) is 1/4, getting 1 head (HT or TH) is 2/4 = 1/2, and getting 2 heads (HH) is 1/4. This probability distribution is called the Probability Mass Function (PMF) because it assigns probabilities to each possible value of X. Since the total probability adds up to 1, it is a valid PMF and helps in understanding the likelihood of different outcomes in the experiment.

In contrast, a continuous random variable can take any value within a given range, such as height, weight, or temperature. Since there are infinitely many possible values, probabilities are assigned over intervals using a probability density function (PDF) instead of individual points.

4.1.2 Discrete Probability Distributions (Probability Mass function)

A **Discrete Probability Distribution (Probability Mass Function (PMF))** describes the probabilities of a **discrete random variable**, which takes a finite or countable number of values. It assigns probabilities to each possible outcome, ensuring that all probabilities are between 0 and 1 and sum to 1.

Let X be a discrete random variable and suppose that the possible values that it can assume are given by $x_1, x_2, x_3, \ldots$, arranged in some order with with probabilities given by

$P(X = x_k) = f(x_k)\ k = 1,2, \ldots (1)$

The **PMF** describes how likely each possible value of **X** is. More generally, the **probability function** or **probability distribution** is written as:

P(X = x) = f(x)

where **x** represents any possible value of **X**. The function **f(x)** must satisfy two conditions:

1. **Non-negativity:** $f(x) \geq 0$ for all values of x.
2. **Total probability equals 1:** $\sum f(x) = 1$.

In cases where **x** is not a possible value of **X**, the function **f(x) = 0**. This ensures that the PMF correctly defines the probability distribution of a discrete random variable.

For example, consider the random experiment of tossing 3 coins. The sample space

$$S = \{(HHH, HHT, HTH, THH, TTT, TTH, THT, HTT\}$$

Let the random variable X be defined as ' Number of heads in each throw'. Then X can assume values 0,1,2,3, with corresponding probabilities

$P[X = 0] = \frac{1}{8} = f(x_1) \geq 0, \quad P[X = 1] = \frac{3}{8} = f(x_2) \geq 0,$

$P[X = 2] = \frac{3}{8} = f(x_3) \geq 0, \quad P[X = 3] = \frac{1}{8} = f(x_4) \geq 0.$

$f(x)$ probability mass function since each probability≥ 0 and sum of the probability

$$= \frac{1}{8} + \frac{3}{8} + \frac{3}{8} + \frac{1}{8} = 1$$

Illustration.4.1.1

Let X be a random variable, and P(X=x) is the PMF given by,

X	0	1	2	3	4	5	6	7
P(X=x)	0	K	2k	2k	3k	k^2	$2k^2$	$7k^2+k$

1. Determine the value of k
2. Find the probability

   (i) $P(X \leq 2)$

   (ii) $P(3 < x \leq 6)$

Solution

1. We know that $\sum P(X = x) = 1$

Therefore, $0 + k + 2k + 2k + 3k + k^2 + 2k^2 + 7k^2 + k = 1$

$$9k + 10k^2 = 1$$

$$10k^2 + 9k - 1 = 0$$

$$10k^2 + 10k - k - 1 = 0$$

$$10k(k + 1) - 1(k + 1) = 0$$

$$(10k - 1)(k + 1) = 0$$

So, $10k - 1 = 0$

and $k + 1 = 0$

$$Therefore, k = \frac{1}{10} \text{ and } k = -1$$

$k = -1$ is not possible because the probability value ranges from 0 to 1. Hence, the value of k $is \frac{1}{10}$.

2. i) Now that we have k = $\frac{1}{10}$, we substitute it into the PMF:

X	0	1	2	3	4	5	6	7
P(X= x)	0	$\frac{1}{10}$(K)	$\frac{2}{10}$(2K)	$\frac{2}{10}$(2K)	$\frac{3}{10}$(3K)	$\frac{1}{100}$(k^2)	$\frac{2}{100}$($2k^2$)	$\frac{7}{100}$+$\frac{1}{10}$($7k^2$+k)

$$P(X \leq 2) = P(X = 0) + P(X = 1) + P(X = 2)$$

$$= 0 + k + 2k$$

$$= 3k$$

$$= \frac{3}{10}$$

ii) $P(3 < x \leq 6) = P(X = 4) + P(X = 5) + P(X = 6)$

$$= 3k + k^2 + 2k^2$$

$$= 3k + 3k^2$$

$$= \frac{3}{10} + \frac{3}{100}$$

$$= \frac{33}{100}$$

Illustration.4.1.2

A random variable X has following probability mass function

X	-2	-1	0	1
P(X)	0.4	K	0.2	0.3

Find k.

Solution

Since P(X) is a probability mass function $\sum P(X = x) = 1$

$$\text{i.e. } 0.4 + k + 0.2 + 0.3 = 1$$

$$k + 0.9 = 1$$

$$k = 1 - 0.9 = 0.1$$

Illustration 4.1.3

If the random variable X takes the values 1,2,3 and 4 such that

$2P(X = 1) = 3P(X = 2) = P(X = 3) = 5P(X = 4)$. Find the probability mass function.

Solution

Let $P(X = 3) = k$, then $2P(X = 1) = k$, $3P(X = 2) = k$,$5P(X = 4) = k$.

$$P(X = 1) = \frac{k}{2}, \quad P(X = 2) = \frac{k}{3}, \quad P(X = 4) = \frac{k}{5}$$

Since P(X) is a probability mass function, $\frac{k}{2} + \frac{k}{3} + k + \frac{k}{5} = 1$

Simplifying $k = \frac{1}{61}$

4.1.3. Binomial Distribution

Let us assume a salesperson is attempting to close deals with potential clients. Each interaction with a client can be viewed as a trial, and the outcome of each trial is either a successful deal (success) or an unsuccessful attempt (failure). Let us denote the probability of successfully closing a deal in any single trial as p, and the probability of failure as $q = 1 - p$.

Now, suppose the salesperson conducts a series of n trials, trying to close deals with different clients independently. The objective is to understand the probability distribution of the number of successful deals (x) among these n trials. This is where the probability function for a binomial distribution comes into play.

The binomial distribution is a probability distribution that models the number of successes in a fixed number of independent trials, where each trial has only two possible outcomes: success or failure.

Key Features:

1. Fixed number of trials (n) – The experiment is repeated n times.
2. Independent trials – The outcome of one trial does not affect another.
3. Two outcomes per trial – Each trial results in either "success" (with probability p) or "failure" (with probability 1−p).
4. Constant probability of success (p) – The probability remains the same for each trial.

The probability mass function (PMF) for a binomial distribution is given by:

$$P(X = x) = {}^{n}C_{x}\, p^{x}\, q^{n-x} \quad \textbf{or} \quad \binom{n}{x}\, p^{x}\, q^{n-x}$$

$$\text{Where } {}^{n}C_{x} = \binom{n}{x} = \frac{n!}{x!(n-x)!}$$

Here, represents the number of ways to choose x successes from n trials,

p^x is the probability of having x successes, and q^{n-x} is the probability of having $n - x$ failures.

For example, let's say the salesperson has a 20% success rate ($p = 0.2$) in closing deals. If the salesperson conducts 10 independent trials, the probability of closing exactly 2 deals ($x = 2$) can be calculated using the binomial distribution PMF. This probability calculation provides insights into the likelihood of achieving a specific number of successful deals in a given number of trials, which is valuable information for sales forecasting and performance evaluation.

Suppose that we have an experiment such as tossing a coin or die repeatedly or choosing a marble from an urn repeatedly. Each toss or selection is called a *trial*. In any single trial there will be a probability associated with a particular event such as head on the coin, 4 on the die, or selection of a red marble. In some cases, this probability will not change from one trial to the next (as in tossing a coin or die). Such trials are then said to be *independent* and are often called *Bernoulli trials* after James Bernoulli who investigated them at the end of the seventeenth century.

Let p be the probability that an event will happen in any single Bernoulli trial (called the *probability of success*). Then $q = 1 - p$ is the probability that the event will fail to happen in any single trial (called the *probability of failure*). The probability that the event will happen exactly x times in n trials (i.e., successes and $n - x$ failures will occur) is given by the probability function

$$f(x) = P(X = x) = \frac{n!}{x!\,(n-x)!} p^x q^{n-x}$$

where the random variable X denotes the number of successes in n trials and $x = 0,1, \dots, n$.

The discrete probability function $P(X = x)$ for the number of successes in n trials, where $x = 0,1, \dots, n$, is commonly referred to as the binomial distribution. This distribution is so named because, for each value of x, it corresponds to the coefficients of the binomial expansion of $(q + p)^n$, where q and p are the probabilities of failure and success, respectively. The binomial expansion is a mathematical expression obtained by raising the binomial $(q + p)$ to the power of n.

The binomial distribution is further illustrated by the binomial expansion formula, which expands $(q + p)^n$ into a sum of terms, each representing the probability of a specific number

$$\binom{n}{x}$$

of successes. The coefficients in the expansion correspond to the number of ways to choose x successes from n trials.

In the special case where $n = 1$, the binomial distribution reduces to the Bernoulli distribution, which represents a single Bernoulli trial. The Bernoulli distribution is characterized by the probability of success (p) and the probability of failure $(q = 1 - p)$ in a single trial, making it a fundamental building block for the broader binomial distribution.

Overall, the binomial distribution is a powerful tool in probability theory and statistics, widely used in various fields, including economics, to model and analyze phenomena involving repeated trials with binary outcomes.

The discrete probability function is often called the *binomial distribution* since for

$x = 0, 1, 2, \dots, n$, it corresponds to successive terms in the *binomial expansion*

$$(q+p)^n = q^n + \binom{n}{1} q^{n-1}p + \binom{n}{2} q^{n-2}p^2 + \cdots + p^n = \sum_{x=0}^{n} \binom{n}{x} p^x q^{n-x}$$

4.1.3.1 Some Properties of The Binomial Distribution

1. **Mean** (μ)

The mean of a binomial distribution is given by $\mu = np$, where n is the number of trials and p is the probability of success in a single trial. This provides the average number of successes expected in n trials.

2. **Variance** (σ^2)

The variance of a binomial distribution is calculated using $\sigma^2 = npq$, where $q = 1 - p$ is the probability of failure.

3. **Standard Deviation** (σ)

The standard deviation is the square root of the variance and is given by $\sigma = \sqrt{npq}$.

Illustration.4.1.4

Seventy-five percent of employed women say their income is essential to support their family. Let *X* be the number in a sample of 200 employed women who will say their income is essential to support their family. What is the mean and standard deviation of *X*

Solution

X is a binomial random variable with $n = 200$ and $p = 0.75(75\% \text{ of employed})$.

The mean is $\mu = np = 200 \times .75 = 150,$

and the standard deviation is $\sigma = \sqrt{npq} = \sqrt{37.5} = 6.12.$

Illustration.4.1.5

A binomial distribution has a mean equal to 8 and a standard deviation equal to 2. Find the values for n and p.

Solution

The following equations must hold: $\mu = 8 = np$

and $\sigma = 4 = npq.$

Substituting 8 for $np,$ in the second equation gives $4 = 8q$, which gives $q = 0.5.$

Since $p + q = 1, p = 1 - 0.5 = 0.5.$

Substituting 0.5 for p in the first equation gives $n\,(.5) = 8$, and it follows that $n = 16.$

Illustration.4.1.6

If the average rainfall is 10 days in every 30 days, obtain the probability that rain will fall on at least 3 days of a given week.

Solution

The probability density function for a binomial distribution is

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

Given $p = \frac{10}{30} = \frac{1}{3}, \quad n = 7, \quad q = 1 - \frac{1}{3} = \frac{2}{3}$

$$P[X \geq 3] = 1 - P[X < 3]$$

$$= 1 - P[X = 0,1,2]$$

$$= 1 - P[X = 0] + P[X = 1] + P[X = 2]$$

$$P(X = 0) = (7\ 0)\left(\frac{1}{3}\right)^0 \left(\frac{2}{3}\right)^{7-0} = \left(\frac{2}{3}\right)^7$$

$$P(X = 1) = (7\ *1)\left(\frac{1}{3}\right)^1 \left(\frac{2}{3}\right)^{7-1} = (7\ 1)\left(\frac{2}{3}\right)^6 \left(\frac{1}{3}\right)$$

$$P(X = 1) = (7 *\ 2)\left(\frac{1}{3}\right)^2 \left(\frac{2}{3}\right)^{7-2} = (7\ 2)\left(\frac{2}{3}\right)^5 \left(\frac{1}{3}\right)^2$$

$$P[X \geq 3] = 1 - 0.5706 = 0.4293$$

Illustration.4.1.7

Ten coins are thrown simultaneously. Find the probability of getting at least seven heads?

Solution

$p =$ Probability of getting a head $= \frac{1}{2}$

q = Probability of not getting a head $= \frac{1}{2}$

The probability of getting x heads in a random throw of 10 coins is

$$P(x) = \frac{10!}{0!\,(10)!} p^x q^{10-x} = 0!\left(\frac{1}{2}\right)^{10}, \quad x = 0,1,2 \ldots .10$$

probability of getting at least seven heads = $P[X \geq 7]$

$$= P[X = 7] + P[X = 8] + P[X = 9] + P[X = 10]$$

$$= \frac{10!}{7!(3)!}\left(\frac{1}{2}\right)^{10} + \frac{10!}{8!(2)!}\left(\frac{1}{2}\right)^{10} + \frac{10!}{7!(1)!}\left(\frac{1}{2}\right)^{10} + \frac{10!}{10!(0)!}\left(\frac{1}{2}\right)^{10}$$

$$= \left(\frac{1}{2}\right)^{10} [120 + 45 + 10 + 1] = \frac{176}{1024} = 0.1699$$

Illustration.4.1.8

A die is tossed 3 times. A success is getting 1 or 6 on a toss. Find the mean and variance of the number of successes.

Solution

Given $n = 3, \quad p = p(getting\ 1\ or\ 6) = \frac{1}{6} = \frac{1}{6} = \frac{1}{3}$

Mean $= np = 3 \times \frac{1}{3} = 1$

Variance $= npq = 3 \times \frac{1}{3} \times \frac{2}{3} = \frac{2}{3}$

Illustration.4.1.9

Suppose that a Central University has to form a committee of 5 members from a list of 20 candidates out of whom 12 are teachers and 8 are students. If the members of the committee are selected at random, what is the probability that the majority of the committee members are students ?

Solution

$p =$ Probability of selecting a student member $= \frac{8}{20} = \frac{2}{5}$

q = Probability of selecting a teacher member $= \frac{12}{20} = \frac{3}{5}$

Let X denote the number of students selected in the committee. Hence, by binomial probability distribution,

$$P(x) = \frac{5!}{x!(5-x)!} p^x q^{5-x} = \frac{5!}{x!(5-x)!} \left(\frac{2}{5}\right)^x \left(\frac{3}{5}\right)^{5-x}, \quad x = 0,1,2,3,4,5.$$

The required probability is given by :

probability of getting at least seven heads = $P[X \geq 3]$

$$= P[X = 3] + P[X = 4] + P[X = 5]$$

$$= \frac{5!}{3!\,2!}\left(\frac{2}{5}\right)^3\left(\frac{3}{5}\right)^{5-3} + \frac{5!}{4!\,(1)!}\left(\frac{2}{5}\right)^4\left(\frac{3}{5}\right)^{5-4} + \frac{5!}{5!\,(0)!}\left(\frac{2}{5}\right)^5\left(\frac{3}{5}\right)^{5-5}$$

$$= \left(\frac{2}{5}\right)^3\left[10 \times \left(\frac{3}{5}\right)^2 + 5 \times \left(\frac{2}{5}\right)\left(\frac{3}{5}\right) + \left(\frac{2}{5}\right)^2\right]$$

$$= \left(\frac{2}{5}\right)^3\left[10 \times \frac{9}{25} + \left(\frac{6}{5}\right) + \left(\frac{4}{25}\right)\right]$$

$$= \left(\frac{2}{5}\right)^3\left[\frac{90 + 30 + 4}{25}\right]$$

$$= \left(\frac{2}{5}\right)^3 \left[\frac{124}{25}\right]$$

$$= \frac{8}{125} \times \frac{124}{25}$$

$$= \frac{992}{3125} = 0.3174$$

4.1.4 Poisson distribution

Poisson distribution was derived in 1837 by a French mathematician Simeon D. Poisson.

Poisson distribution may be obtained as a limiting case of Binomial probability distribution under the following conditions:

(i) n, the number of trials is indefinitely large i.e., $n \to \infty$.
(ii) p, the constant probability of success for each trial is infinitely small i.e., $p \to 0$.

(iii) $np = \lambda$, (say), is finite.

Under the above three conditions the Binomial probability function tends to the probability function of the Poisson distribution given by

$$p(x) = P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}, x = 0, 1, 2, 3, \ldots$$

where X is the number of successes (occurrences of the event), $\lambda = np$.

4.1.4.1 Some Properties of The Poisson Distribution

1. **Mean** (λ)

The mean of a poisson distribution is given by $\lambda = np$, where n is the number of trials and p is the probability of success in a single trial. This provides the average number of successes expected in n trials.

2. **Variance** (λ)

The variance of a binomial distribution is calculated using $\sigma^2 = \lambda$,

3. **Standard Deviation** (σ)

The standard deviation is the square root of the variance and is given by $\sigma = \sqrt{\lambda}$.

Illustration 4.1.10

If 5% of the electric bulbs manufactured by a company are defective, use Poisson distribution to find the probability that in a sample of 100 bulbs

(i) none is defective,
(ii) 5 bulbs will be defective. (Given : $e^{-5} = 0 \cdot 007$)

Solution

Given $n = 100$,

p =probability of defective bulbs $= 5\% = \frac{5}{100}$

$\lambda = np = 100 \times \frac{5}{100} = 5$

$$P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!} \ , x = 0, 1, 2, 3, \ldots$$

i) $P(X = 0) = \frac{e^{-5}5^0}{0!} = e^{-5} = 0.007$

ii) $P(X = 5) = \frac{e^{-5}5^5}{5!} = \frac{0.007 * 3125}{120} = 0.182$

Illustration.4.1.11

A manufacturer of cotter pins knows that 5% of his product is defective. If he sells cotter pins in boxes of 100 and guarantees that not more than 10 pins will be defective, what is the approximate probability that a box will fail to meet the guaranteed quality?

Solution

We are given-$n = 100$.

p - Probability of a defective pin $= 5\% = \frac{5}{100} = 0.05$

λ = Mean' number of defective pins in a box of 100

$= np = 100 * 0 \cdot 05$

$= 5$

Since 'p' is small, we may use Poisson distribution.

Probability of x detective pins in a box of 100 is

$$P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!} = \frac{e^{-5}\ 5^x}{x!} \ , x = 0, 1, 2, 3, \ldots$$

$$P[X > 10] = 1 - P[X \leq 10\ )$$

$$= 1 - \{P[X = 0] + P[X = 1] + \cdots + P[X = 10]$$

$$= 1 - \{\frac{e^{-5}\ 5^0}{0!} + \frac{e^{-5}\ 5^1}{1!} + \cdots + \frac{e^{-5}\ 5^{10}}{10!}\}$$

$$= 1 - e^{-5}\left\{1 + 5 + \cdots + \frac{5^{10}}{10!}\right\} = 1 - 0.999 = 0.001$$

The probability that a box will fail to meet the guaranteed quality is **0.001** (or **0.1%**).

Illustration 4.1.12

The number of accidents in a year to taxi drivers in a city follows a Poisson distribution with mean 3. Out of 2000 taxi drivers , find the number of drivers with more than 3 accidents in a year?

Solution

We are given- $n = 2000, \lambda = 3$

$$P[X > 3] = 1 - P[X \leq 3\ )$$

$$= 1 - \{P[X = 0] + P[X = 1] + P[X = 2] + P[X = 3]$$

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

$$= 1 - e^{-3}\{1 + 3 + \frac{9}{2} + \frac{27}{6}\}$$

$$= 1 - e^{-3}\{1 + 3 + \frac{9}{2} + \frac{27}{6}\}$$

$$= 1 - e^{-3}\{13\}$$

$$= 1\text{- } 13 \times 0.0498 \qquad [e^{-3}=0.0498]$$

$$= 1 - 0.6474 = 0.3528$$

Compute Expected Number of Drivers

Out of 2000 drivers, the expected number of drivers with more than 3 accidents is:

$0.3528 \times 2000 = 705.6$. Since the number of drivers must be a whole number, we round it to 706.

Illustration.4.1.13

Suppose that on an average one house in 1000 in a certain district has a fire during a year. If there are 2000 houses in that district what is the probability that exactly 5 houses will have a fire during the year?

Solution

given-$n = 2000,\ p = \frac{1}{1000}$

$$\lambda = np = 2000 \times \frac{1}{1000} = 2$$

$$P[X = 5] = \frac{e^{-2}2^5}{5!}$$

$e^{-2} = 0.1353$

$$\frac{e^{-2}1^5}{5!} = \frac{0.1353 \times 32}{120} = 0.0361$$

Illustration.4.1.14

A car hire firm has 2 cars which it hires out day by day. The number of demands for a car on each day is distributed as Piosson distribution with mean 1.5. Calculate the proportion of days on which 1) there is no demand 2) Some demand is refused.

Solution

Given $\lambda = 1.5$

1. proportion of days with no demand $= P(0) = \frac{e^{-1.5} \times 1.5^0}{0!} = 0.2231$

   Proportion of days with no demand = **22.31%** (or **0.2231**).

2. proportion of days on which some demand is refused=$P[X > 2]$

$$= 1 - \{P[X = 0] + P[X = 1] + P[X = 2]\}$$

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

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

$$= 0.191$$

So, the proportion of days where some demand is refused is 19.12%.

Recap

- Discrete random Variables: Discrete random variables are random variables that take on distinct, separate values with gaps between them, often associated with counting and finite outcomes in probability distributions.
- Binomial Distribution: The binomial distribution is a discrete probability distribution that models the number of successes in a fixed number of independent and identical Bernoulli trials, characterized by two parameters: the probability of success and the number of trials.
- Poisson distribution: A special type of Binomial Distribution in which the number of occurrences of an event is very large and probability of success is very small, it is a distribution of rare events.

Objective Questions

1. Give an example of a discrete random variable?
2. A probability mass function (PMF) is used for which type of random variable?
3. A binomial distribution is characterized by how many parameters?
4. In a binomial distribution, what does "independent trials" mean?

5. What is the key assumption in a binomial distribution?
6. If a binomial experiment consists of 5 trials with a success probability of 0.4, what is the probability of exactly 2 successes? (Use the binomial formula)
7. What is a discrete probability distribution?
8. What is the sum of probabilities of all outcomes in a discrete probability distribution
9. There are only two possible outcomes in each trial (success or failure) which distribution is applicable
10. What is the mean of a binomial distribution with n trials and probability of success p
11. If n=10 and p=0.4, then what is the variance of the binomial distribution
12. When the number of occurrences of an event is very large, and probability of success is very small, which distribution is applicable?

Answer

1. The number of heads obtained in 10 coin tosses
2. Discrete
3. 2 (Number of trials n and probability of success p)
4. The outcome of one trial does not affect the next
5. Each trial results in exactly one of two outcomes (success or failure)
6. 0.3456
7. The probability of all possible outcomes of a discrete random variable
8. Equal to 1
9. A binomial distribution
10. np
11. 2.4
12. The Poisson distribution

Assignment Questions

1. Define a discrete random variable and provide three real-life examples.
2. A factory produces electronic chips with a 5% defect rate. If a batch contains 10 chips, find the probability that:

a. Exactly 2 chips are defective.

b. At most 2 chips are defective.

3. The number of printing errors in a book follows a Poisson Distribution with a mean of 2 errors per 100 pages. What is the probability that a 150-page book contains exactly 3 errors?
4. Thirty percent of the trees in a forest are infested with a parasite. Fifty trees are randomly selected from this forest and X is defined to equal the number of trees in the 50 sampled that are infested with the parasite. The infestation is uniformly spread throughout the forest. Identify the values for n, p, and q. Suppose we define Y to be the number of trees in the 50 sampled that are not infested with the parasite. Then Y is a binomial random variable.

a. What are the values of n,p, and q for Y ?

b. The event X=20 is equivalent to the event that Y=a. Find the value for a.

5. The mean of a binomial distribution is 4 and its standard deviation is $\sqrt{3}$. What are the values of n, p and q with usual notations ?
6. In a Binomial distribution with 6 independent trials, the probabilities of 3 and 4 successes are found to be 0·2457 and 0·0819 respectively. Find the parameter 'p' of the Binomial distribution.

7. A manufacturer of blades knows that 5% of his product is defective. If he sells blades in boxes of 100, and guarantees that not more than 10 blades will be defective, what is the probability (approximately) that a box will fail to meet the guaranteed quality ?
8. In a certain factory turning out optical lenses, there is a small chance 1/500 for any one lens to be defective. The lenses are supplied, in packets of 10. Use Poisson distribution to calculate the approximate number of packets containing no defective, one defective, two defective, three defective lenses respectively in a consignment of 20,000 packets. You are given that $e^{-0.02} = 0 \cdot 9802$.

Suggested reading

1. John Wiley & Sons, Inc., NY, USA, 2010 3. Montgomery, D. C. and G. C. Runger. Applied Statistics and Probability for Engineers. 5th Edition. John Wiley & Sons, Inc., NY, USA, 2011.
2. Kolman, Busby, Ross and Rehman, Discrete Mathematical Structures for Computer Science, Pearson Education,6th Edition, 2017.
3. Erwin Kreyszig, Advanced Engineering Mathematics, Wiley India, 9th Edition, 2011.

References

1. Montgomery, D. C., & Runger, G. C. (2011). *Applied statistics and probability for engineers* (5th ed.). John Wiley & Sons, Inc.
2. Kolman, B., Busby, R. C., Ross, S. T., & Rehman, M. (2017). *Discrete mathematical structures for computer science* (6th ed.). Pearson Education.
3. Kreyszig, E. (2011). *Advanced engineering mathematics* (9th ed.). Wiley India.
4. Ross, S. (2019). *A first course in probability* (9th ed.). Pearson Education.
5. Durrett, R. (2019). *Probability: Theory and examples* (5th ed.). Cambridge University Press.

UNIT- 2

Continuous probability distributions- Normal, Exponential and Uniform Distributions

Learning outcome

At the end of this unit, the student will be able to:

- State the properties of Normal, Exponential, and Uniform distributions.
- Understand continuous Probability Distributions:
- Exploring Normal, exponential and uniform Distributions:
- Recognize the differences between Normal, Exponential, and Uniform distributions from given graphical representations.

Pre-requisites

In weather forecasting, understanding random variables, outcomes, and events helps predict temperature variations (which often follow a Normal Distribution). Familiarity with descriptive statistics is crucial for interpreting distribution parameters—such as the mean and standard deviation in quality control, where manufacturers ensure product weights remain within acceptable limits (following a Normal Distribution).

Additionally, a basic understanding of calculus, particularly integration, is essential when dealing with continuous distributions. For instance, in customer service, the Exponential Distribution helps model the time between customer arrivals at a call center. Similarly, the Uniform Distribution applies in random number generation, such as lottery draws, where each number has an equal chance of selection. By mastering these foundational concepts, one can effectively apply probability distributions to real-world problems in science, business, and engineering.

Keywords:

Normal Distribution, exponential Distributions, uniform Distributions

Discussion

4.2.1. Continuous probability function

Random variables can take on different forms based on the nature of the outcomes they represent. A discrete random variable is one that can assume a finite or countably infinite set of distinct values. For instance, the number of students in a classroom or the count of defective items in a production batch are examples of discrete random variables. On the other hand, a non discrete random variable takes on an uncountable infinite number of values and often corresponds to continuous phenomena, such as the height of individuals or the temperature in a given location.

A function $f(x)$ that satisfies the above requirements is called a *probability function* or *probability distribution* for a continuous random variable, but it is more often called a *probability density function* or *simply density function*. Any function $f(x)$ satisfying Properties 1 and 2 above will automatically be a density function.

A random variable X is said to be absolutely continuous, or simply continuous, if the Probability density function is

1. $f_i(x_i) \geq 0$
2. $\int_{-\infty}^{\infty} f_i(x_i)\, dx_i = 1$

Illustration.4.2.1

Find the value of k if $f(x) = k(2-x)\ \ 0 < x < 2$ is a Probability density function.

Solution

Since $f(x)$ is a continuous probability density function $\int_0^2 \quad f(x)\, dx = 1$

$$\int_0^2 \quad k(2-x)\, dx = 1$$

$$k\left(2x - \frac{x^2}{2}\right)_0^2 = 1$$

$$k\left(4 - \frac{4}{2}\right) = 1$$

$$k(4-2) = 1$$

$$k = \frac{1}{2}$$

Illustration.4.2.2

Show that $f(x) = \frac{x+1}{2} \quad |x| < 1$

$= 0$ elsewhere represents the Probability density function of a random variable X.

Solution

If $f(x)$ is a Probability density function then $\int f(x)\,dx = 1$

$$\text{Consider} \quad \int_{-1}^{1} \frac{x+1}{2} = \frac{1}{2}\int_{-1}^{1} x + 1 \quad dx$$

$$= \frac{1}{2}\left(\frac{x^2}{2} + x\right)_{-1}^{1}$$

$$= \frac{1}{2}\left(\frac{1}{2} + 1 - (\frac{1}{2} - 1)\right)_{-1}^{1}$$

$$= \frac{1}{2} \times 2$$

$$= 1$$

4.2.2. Normal Distribution

One of the most important examples of a continuous probability distribution is the *normal distribution*, sometimes called the *Gaussian distribution*. The density function for this distribution is given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} \qquad -\infty < x < \infty$$

where μ and σ are the mean and standard deviation, respectively. The corresponding distribution function is given by

$$F(x) = P(X \leq x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-(v-\mu)^2/2\sigma^2} dv$$

If X has the distribution function, we say that the random variable X is *normally distributed* with mean μ and variance σ^2.

If we let Z be the standardized variable corresponding to X, i.e., if we let

$$z = \frac{x - \mu}{\sigma} \approx N(0,1)$$

then the mean or expected value of Z is 0 and the variance is 1. In such cases the density function for Z can be formally placing $\mu = 0$ and $\sigma = 1$, yielding

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$$

This is often referred to as the *standard normal density function*.

A graph of the density function, sometimes called the *standard normal curve*. In this graph, in figure 4.2.1 we have indicated the areas within 1, 2, and 3 standard deviations of the mean (i.e., between $z = -1$ and $+1$, $z = -2$ and $+2, z = -3$ and +3) as equal, respectively, to 68.27%, 95.45% and 99.73% of the total area, which is one. This means that

$$P(-1 \leq Z \leq 1) = 0.6827 \;\; P(-2 \leq Z \leq 2) = 0.9545 \;\; P(-3 \leq Z \leq 3) = 0.9973$$

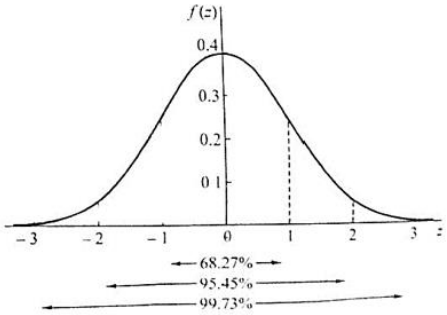


Fig 4.2.1 standard normal curve

Illustration.4.2.3

Express the areas shown in the following two standard normal curves, figure 4.2.2 as a probability statement and find the area of each one.

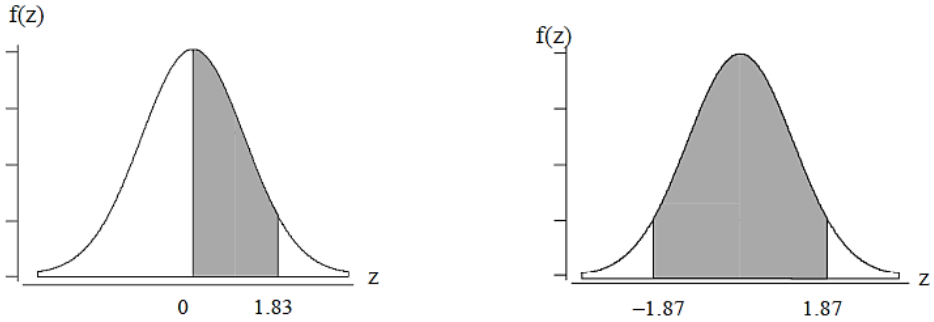


Fig 4.2.2 standard normal curve

Solution

The area under the curve on the left is represented as $P(0 < Z < 1.83)$ and from the standard normal distribution table is equal to 0.4664. The area under the curve on the right is represented as $P(-1.87 < Z < 1.87)$ and from the standard normal distribution table is $2 \times .4693 = 0.9386$.

Illustration.4.2.4

The distribution of complaints per week per 100,000 passengers for all airlines in a country is normally distributed with $\mu = 4.5$ and $\sigma = 0.8$. Find the standardized values for the following observed values of the number of complaints per week per 100,000 passengers: (a) 6.3; (b) 2.5 ; (c) 4.5 ; (d) 8.0 .

Solution

(a) The standardized value for 6.3 is found by $z = \frac{x-\mu}{\sigma} = \frac{6.3-4.5}{.8} = 2.25.$

(b) The standardized value for 2.5 is found by $z = \frac{x-\mu}{\sigma} = \frac{2.5-4.5}{.8} = -2.50.$

(c) The standardized value for 4.5 is found by $z = \frac{x-\mu}{\sigma} = \frac{4.5-4.5}{.8} = 0.00.$

(d) The standardized value for 8.0 is found by $z = \frac{x-\mu}{\sigma} = \frac{8.0-4.5}{.8} = 4.38.$

Illustration.4.2.5

The net worth of senior citizens is normally distributed with mean equal to $\$225{,}000$ and standard deviation equal to $\$35{,}000$. What percent of senior citizens have a net worth less than $\$300{,}000$?

Solution

Let X represent the net worth of senior citizens in thousands of dollars. The percent of senior citizens with a net worth less than $\$300{,}000$ is found by multiplying $P(X < 300)$ times 100. The probability $P(X < 300)$ is shown in figure below.

The event $X < 300$ is equivalent to the event $Z < \frac{300-225}{35} = 2.14$. The probability that

$Z < 2.14$ is represented as the shaded area in Fig. 4.4.3 The probability that Z is less than 2.14 is found by adding $P(0 < Z < 2.14)$ to 0.5, which equals $0.5 + 0.4838 = 0.9838$. We can conclude that 98.38% of the senior citizens have net worths less than $\$300{,}000$.

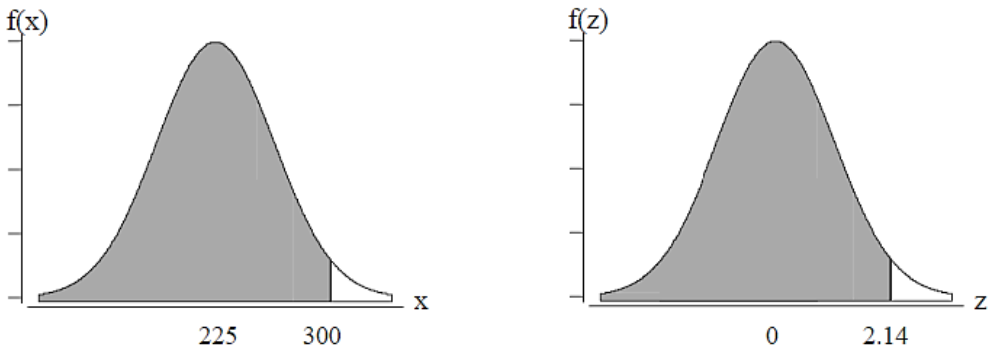


Fig 4.2.3 standard normal curve

Illustration 4.2.6

The average test marks in a particular class is 79 and standard deviation is 5. If the marks are normally distributed how many students in a class of 200 did not receive marks between 75 and 82?

Solution

Given $\mu = 79,\ \sigma = 5,\ n = 200$

$$z = \frac{x - \mu}{\sigma} = \frac{75 - 79}{5} = -0.8$$

$$z = \frac{x - \mu}{\sigma} = \frac{82 - 79}{5} = 0.6$$

$$P[75 < X < 82] = P[-0.8 < Z < 0.6]$$

$$= 0.2881 + 0.2257 = 0.5138$$

$$1 - P[75 < X < 82] = 1 - 0.5138 = 0.4862$$

Number of students $= \ 0.4862 \times 200 = 97.24$

Since the number of students must be an integer, we round to **97**.

4.2.3. Exponential Distribution

A continuous random variable X as with non-negative values is said to have an exponential distribution with parameter > 0, if its probability density function is given by

$$f(x) = \lambda\, e^{-\lambda x} \quad if\ \lambda > 0$$

$$= 0 \quad if\ \lambda \leq 0$$

4.2.4 Some Properties of The Exponential Distribution

1. Mean

The mean of an e exponential distribution is given by $\frac{1}{\lambda}$ where λ is the parameter.

2. Variance

The variance of a exponential distribution is $\frac{1}{\lambda^2}$.

Illustration.4.2.7

The amount of time that a watch will run without having to be reset is a random variable having an exponential distribution with mean 120 days. Find the probability that such a watch will

i) have to reset in less than 24 hours

ii) not have to reset in at least 180 days?

Solution

Given mean $\frac{1}{\lambda} = 120$

The probability density function is given by

$$f(x) = \lambda\, e^{-\lambda x} \;\; if\; \lambda > 0$$

$$= \frac{1}{120}\, e^{-\frac{1}{120}x}$$

i) P [watch have to be reset in less than 24 hours]

= P[watch will run for less than 24 days]

$= P[X < 24]$

$$= \int_0^{24} f(x)\, dx$$

$$= \int_0^{24} \frac{1}{120}\, e^{-\frac{1}{120}x} dx$$

$$= \frac{1}{120} \int_0^{24} e^{-\frac{1}{120}x} dx$$

$$= \frac{1}{120} \left(\frac{e^{-\frac{1}{120}x}}{-\frac{1}{120}} \right)_0^{24}$$

$$= 1 - \left(e^{-\frac{24}{120}} - e^{0} \right) = 1 - \left(e^{-\frac{24}{120}} - 1 \right) = 1 - 0.9917 = 0.00830$$

Probability that the watch has to be reset in less than 24 hours: 0.0083 (0.83%)

ii. Probability that the watch will not have to be reset for at least 180 days

$$P(T \geq 180) = 1 - P(T < 180)$$

$$= e^{-\lambda(180)}$$

Substituting $\lambda = \frac{1}{120}$

$$P(T \geq 180) = e^{-\frac{120}{180}} = e^{-1.5}$$

Approximating $e^{-1.5}$

$e^{-1.5} \approx 0.2231$

Probability that the watch will not have to be reset in at least 180 days: 0.2231 (22.31%)

Illustration.4.2.8

The mileage which a car owner gets with a certain kind of tyre is a random variable having exponential distribution with mean 40,000 kms. Find the probability that one of these tyres will last i) at least 30000 kms 2) atmost 35,000 kms.

Solution

Given mean $\frac{1}{\lambda} = 40{,}000$

The probability density function is given by

$$f(x) = \lambda\, e^{-\lambda x} \;\; if\; \lambda > 0$$

$$= \frac{1}{40000}\; e^{-\frac{1}{40000}x}$$

i) P [one of these tyres will last at least 30000 kms]

$$= P[X > 30000]$$

$$= \int_{30000}^{\infty} \frac{1}{40000}\; e^{-\frac{1}{40000}x}\, dx$$

$$= \frac{1}{40000}\int_{30000}^{\infty} e^{-\frac{1}{40000}x}\, dx$$

$$= \frac{1}{40000}\left(\frac{e^{-\frac{1}{40000}x}}{-\frac{1}{40000}}\right)_{30000}^{\infty}$$

$$= e^{-\frac{30000}{40000}} = 0.4724$$

Thus, the probability that the tyre lasts at least 30,000 km is 0.4724 (or 47.24%).

ii. Probability that the tyre lasts at most 35,000 km

We need to compute:

$$P(X \leq 35000) = 1 - e^{-\lambda(35000)}$$

Substituting $\lambda = \frac{1}{4000}$

$$P(X \leq 35000) = 1 - e^{-\frac{35000}{40000}}$$

Using $e^{-0.875} \approx 0.4169$

$$P(X \leq 35000) = 1 - 0.4169 = 0.5831$$

Thus, the probability that the tyre lasts at most 35,000 km is 0.5831 (or 58.31%).

Illustration.4.2.9

The time in hours required to repair a machine is exponentially distributed with $\lambda = \frac{1}{20}$. What is the probability that the required time

i) Exceeds 30 hrs.
ii) In between 16 hrs. and 24 hrs.
iii) Atmost 10 hrs

Solution

$$f(x) = \lambda\, e^{-\lambda x} \;\; if\; \lambda > 0$$

$$= \frac{1}{20}\, e^{-\frac{1}{20}x} \quad x > 0$$

i) Probability that the required time exceeds 30 hours

We need to compute:

$$P(X>30)=e^{-\lambda(30)}$$

Substituting λ=120

$$P(X>30) = e^{-120\times 30} = e^{-3600}$$

Since e−3600 is an extremely small number, it is **approximately 0**.

Thus, **P(X>30) ≈ 0**.

Probability that repair time exceeds 30 hours: ≈0

ii) $P[16 < X < 24]$

$$= \int_{16}^{24} f(x)\, dx$$

$$= \int_{16}^{24} \frac{1}{20}\, e^{-\frac{1}{20}x} dx$$

$$= \frac{1}{20}\left(\frac{e^{-\frac{1}{20}x}}{-\frac{1}{20}}\right)_{16}^{24}$$

$$= \frac{1}{20}\left(e^{-\left(\frac{24}{20}\right)} - e^{-\left(\frac{16}{20}\right)}\right) = 0.148$$

Probability that repair time is between 16 and 24 hours: ≈ 0

iii) We need to compute:

$$P(X\leq 10) = 1-e^{-\lambda(10)}$$

Substituting λ=120:

$$P(X \leq 10) = 1 - e - 120 \times 10$$

$$= 1 - e^{-1200}$$

Since e^{-1200} is an extremely small number, we get:

$$P(X \leq 10) \approx 1$$

Thus, $\mathbf{P(X \leq 10) \approx 1}$

Probability that repair time is at most 10 hours: ≈1

4.2.4 Uniform Distribution

Consider the random variable X representing the flight time of an aeroplane travelling from Trivandrum to New Delhi. Suppose the flight time can be any value in the interval from 120 minutes to 140 minutes. Because the random variable X can assume any value in that interval, X is a continuous rather than a discrete random variable. Let us assume that sufficient actual flight data are available to conclude that the probability of a flight time within anyone-minute interval is the same as the probability of a flight time within any other one-minute interval contained in the larger interval from 120 to 140 minutes.With everyone-minute interval being equally likely, the random variable X is said to have a uniform probability distribution.

If x is any number lying in the range that the random variable X can take then the probability density function, which defines the uniform distribution for the flight-time random variable, is:

$$f(x) = \frac{1}{20} \qquad for\ 120 \leq x \leq 140$$

$$= 0 \qquad \text{elsewhere}$$

The probability function for uniform distribution is

$$f(x) = \frac{1}{b-a} \qquad for\ a \leq x \leq b$$

$$= 0 \qquad \text{elsewhere}$$

4.2.4.1 Properties of Uniform Distribution

1. **Mean**

The mean of a uniform distribution is $\frac{a+b}{2}$, where x is defined for $a \leq x \leq b$

2. **Variance**

The variance of a uniform distribution is $\frac{(a-b)^2}{12}$.

Illustration.4.2.10

A bus arrives every 15 minutes at a bus stop. Assuming that the waiting time X for the bus is Uniformly distributed, find the probability that a person has to wait for the bus 1) more than 10 minutes 2) between 5 and 10 minutes.

Solution

$$\text{Given } f(x) = \frac{1}{15} \quad 0 < x < 15$$
$$= 0, \quad otherwise$$

1) $P[X > 10] = \int_{10}^{15} \frac{1}{15}\ dx = \frac{1}{15}\int_{10}^{15} dx = \frac{1}{15}(15-10) = \frac{5}{15} = \frac{1}{3}$
2) $P[5 < X < 10] = \int_{5}^{10} \frac{1}{15}\ dx = \frac{1}{15}\int_{5}^{10} dx = \frac{1}{15}(10-5) = \frac{5}{15} = \frac{1}{3}$

Illustration.4.2.11

Subway trains on a certain line run every half hour between midnight and six in the morning. What is the probability that a man entering the station at a random time during this period will have to wait at least 20 minutes?

Solution

Let the random variable X denote the waiting time (in minutes) for the next train. Under the assumption that the man enters the station at a **random time** between midnight and 6 AM. Let X be the waiting time for the next train, which is uniformly distributed between 0 and 30 minutes (since the trains run every half hour).

Thus, $X \sim U(0,30)$, meaning the probability density function (PDF) is:

$$f(x) = \frac{1}{30} \qquad 0 \leq x \leq 30$$

The probability that he has to wait at least 20 minutes is

$$[X > 20] = \int_{20}^{30} \frac{1}{30}\ dx = \frac{1}{30}\int_{20}^{30} dx = \frac{1}{30}(30-20) = \frac{10}{30} = \frac{1}{3}$$

Illustration 4.2.12

Subway trains on a certain line run every half hour between midnight and six in the morning. What is the probability that a man entering the station at a random time during this period will have to wait at least 20 minutes?

What is the probability that a passenger will wait more than 3 minutes for a shuttle train

Solution

Let the random variable X denote the waiting time (in minutes) for the shuttle train.

X is distributed uniformly on (0, 5), with probability function

$$f(x) = \frac{1}{5}$$

The probability that he has to wait more than 3 minutes is

$$[X > 3] = \int_3^5 \frac{1}{5}\, dx = \frac{1}{5}\int_3^5 dx = \frac{1}{5}(5-3) = \frac{2}{5}$$

Recap

- Continuous Variables: Continuous variables are random variables that can take any value within a range, often associated with measurements and infinite possible outcomes, characterized by a continuous probability distribution.
- Normal Distribution: The normal distribution, also known as the Gaussian distribution or bell curve, is a continuous probability distribution characterized by a symmetric, bell-shaped curve, where the majority of observations cluster around the mean.
- Exponential Distribution: The exponential distribution is a continuous probability distribution often used to model the time between events in a Poisson process. It describes situations where events occur independently and at a constant average rate
- Uniform distribution is a type of probability distribution where all outcomes within a specified range are equally likely.

Objective Questions

1. A normal distribution is also known as
2. The mean, median, and mode of a normal distribution are:
3. The probability density function (PDF) of a normal distribution is:
4. The exponential distribution is used to model
5. If the average time between customer arrivals is 5 minutes, the rate parameter λ for the exponential distribution is:
6. Which distribution is used when all outcomes within a specified range are equally likely.
7. Give an example of a uniform distribution?

Anseres

1. Gaussian distribution
2. All equal
3. Symmetrical about the mean
4. The time between events in a Poisson process
5. 1/5
6. Uniform distribution
7. Rolling a fair die

Assignment

1. The hospital cost for individuals involved in accidents who do not wear seat belts is normally distributed with mean $Rs.7500$ and standard deviation $Rs.1200$.

   (a) Find the cost for an individual whose standardized value is 2.5.

   (b) Find the cost for an individual whose bill is 3 standard deviations below the average.

2. The average TV-viewing time per week for children ages 2 to 11 is 22.5 hours and the standard deviation is 5.5 hours. Assuming the viewing times are normally distributed, find the following.

a) What percent of the children have viewing times less than 10 hours per week?
b) What percent of the children have viewing times between 15 and 25 hours per week?
c) What percent of the children have viewing times greater than 40 hours per week?

3. In a certain examination 15% of the candidates passed with distinction while 25% of them failed. It is known that a candidate fails if he obtains less than 40 marks (out of 100) while he must obtain at lest 75 marks in order to pass with distinction. Determine mean and standard deviation of the distribution of marks assuming this to be normal.
4. If the cauliflowers on a truck are classified as A, B and C according to a size-weight index as : under 75, between 75 and 80, and above 80; find approximately (assuming a normal distribution) the mean and standard deviation of a lot in which A are 58%, B are 38% and C are 4%.
5. The time in hours required to repair a machine is exponentially distributed with $\lambda = \frac{1}{30}$. What is the probability that the required time

   Exceeds 20 hrs. In between 18 hrs. and 22 hrs. Atmost 12 hrs.

6. Assume the weight of a randomly chosen American passenger car is a uniformly distributed random variable ranging from 2,500 pounds to 4,500 pounds.

   a. What is the mean and standard deviation of weight of a randomly chosen vehicle?

   b. What is the probability that a vehicle will weigh less than 3,000 pounds?

   c. More than 3,900 pounds?

   d. Between 3,000 and 3,800 pounds?

Suggested reading

1. John Wiley & Sons, Inc., NY, USA, 2010 3. Montgomery, D. C. and G. C. Runger. Applied Statistics and Probability for Engineers. 5th Edition. John Wiley & Sons, Inc., NY, USA, 2011.
2. Kolman, Busby, Ross and Rehman, Discrete Mathematical Structures for Computer Science, Pearson Education,6th Edition, 2017.
3. Erwin Kreyszig, Advanced Engineering Mathematics, Wiley India, 9th Edition, 2011.

References

1. K Huffman and R Kunze, Linear Algebra, Pearson Education, 2nd Edition, 2005.
2. Thomas M. Apostol, Calculus, Wiley, 2nd Edition, 1991 ISBN 960-07-0067-2.
3. Michael Spivak. Calculus, publish or Perish, 2008, ISBN 978-0914098911.
4. Ross L. Finney, Maurice D.Weir. and Frank R. Giordano. Thomas's Calculus, Pearson l2th Edition 2009.
5. Bendat, J. S. and A. G. Piersol. Random Data: Analysis and Measurement Procedures. 4th Edition.

UNIT- 3

Central limit theorem and its applications

Learning outcome

At the end of this unit, the student will be able to:

- Understanding the Normal Approximation of Sample Means
- Define the Central Limit Theorem (CLT) and explain its significance in probability and statistics.
- Explain the importance of CLT in statistical inference
- Apply Central limit theorem in real life problems

Pre-requisites

A strong foundation in probability theory, random variables, and statistical distributions is essential for understanding the Central Limit Theorem (CLT) and its applications. Knowledge of concepts like mean, variance, and standard deviation helps in grasping how sample means behave as sample size increases. Familiarity with different types of distributions, particularly normal and non-normal distributions, is important to see how CLT applies regardless of the population shape. Sampling methods and the law of large numbers are also crucial, as they explain why larger samples yield more reliable estimates. CLT has widespread applications in fields like survey analysis, where it helps estimate population parameters; quality control, where it ensures product consistency through sample-based checks; and finance, where it aids in risk assessment and portfolio analysis by approximating distributions of returns.

Keywords

Continuous probability function, Central limit theorem, c

Discussion

4.3.1 Continuous probability function

The normal distribution is not confined to just making approximations to the binomial distribution. In general, the normal curve results whenever there are large numbers of independent small factors influencing the final outcome. It is for this reason that many practical distributions, be it the distribution of annual rainfall, weight of birth babies, the daily attendance in the out-patient department of a large hospital are more or less normal, if sufficient large numbers of items are included in the population.

The significance of a normal curve is much more than this. It can be shown that even when the original population is not normal, if we draw samples of n items from it and obtain the distribution of sample means become more and more normal as the sample size n increases. This fact can be proved mathematically, and is stated as the Central Limit Theorem.

If we take sample of size n from any population and calculate $\underline{x}$, the sample distribution of $\underline{x}$,

Will approach the normal distribution as the sample size n increases.

By understanding the intricacies of the Central Limit Theorem, we gain an invaluable tool that transcends specific datasets and contexts. This theorem empowers us to interpret and make predictions about the distribution of sums, averages, and aggregates in scenarios where we may not have explicit knowledge about the underlying distribution of individual variables. It is a cornerstone of modern statistical thinking, offering us a lens through which to understand the patterns that emerge from the accumulation of random variables in diverse and complex scenarios.

4.3.1.1 Standard error

The standard deviation of the sampling distribution is known as standard error (SE). The reciprocal of the SE of a statistic gives a measure of the precision or reliability of the estimate of the parameter. The concept of SE is extremely useful in the testing of statistical hypothesis and estimation of parameters.

4.3.2 Central Limit Theorem

The Central Limit Theorem is stated as follows:

Let $X_1, X_2, \ldots, X_n$ be a sample from a population having mean μ and standard deviation σ. For *n* large, the sum $S_n = X_1 + X_2 + \cdots + X_n$ will approximately have a normal distribution with mean $n\mu$ and standard deviation $\sqrt{n}\,\sigma$.

$$z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}} = \frac{S_n - n\mu}{SE}$$

will approximately have a standard normal distribution with mean $n\mu$ and standard deviation $\sqrt{n}\,\sigma$.

OR

Let $X_1, X_2, \ldots, X_n$ be a sample from a population having mean μ and standard deviation σ. For *n* large, and if $\underline{x} = \sum \frac{X_i}{n}$ then $z = \frac{x-\mu}{\sigma/\sqrt{n}}$ will approximately have a standard normal distribution with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$.

This theorem stands as a cornerstone concept, with far-reaching implications that significantly shape our understanding of probability distributions and statistical behaviour.

4.3.1.Illustration

The number of pieces of mail that a dependent receives each day can be modelled by a distribution having mean 44 and standard deviation 8. For a random sample of 25 days, what can be said about the probability that the sample mean will be less than 40 or greater than 48, using Central Limit Theorem

Solution

Given Information:

- The number of pieces of mail received per day follows a distribution with:
  - Mean μ=44
  - Standard deviation σ=8
- Sample size n=25
- We need to find $P(\bar{X}<40)+P(\bar{X}>48)$using the Central Limit Theorem (CLT).

According to the Central Limit Theorem (CLT), for large n, the sampling distribution of the sample mean $\bar{X}$ is approximately normal with:

- Mean of sample mean:

$$\mu_{\bar{X}}=\mu=44$$

- Standard deviation of sample mean (Standard Error, SE):$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{8}{\sqrt{25}} = \frac{8}{5} = 1.6$

Convert to Z-scores

We calculate the Z-scores for $\bar{X}=40$ and $\bar{X}=48$

For $\bar{X}=40$:

$$Z = \frac{40-44}{1.6} = \frac{-4}{1.6} = -2.5$$

For $\bar{X}$=48:

$$Z = \frac{48-44}{1.6} = \frac{4}{1.6} = 2.5$$

Find Probabilities from Z-table

From the standard normal table:

- $P(Z<-2.5) = 0.0062$
- P(Z>2.5)=0.0062

Since we need:

$$P(\bar{X}<40)+P(\bar{X}>48)$$

$$= P(Z<-2.5)+P(Z>2.5)$$

$$= 0.0062+ 0.0062 = 0.0124$$

The probability that the sample mean will be less than 40 or greater than 48 is 0.0124 (or 1.24%).

Illustration 4.3.2

An electrical firm manufactures light bulbs that have length of life that is approximately normally distributed with mean 800 hrs and standard deviation 40 hrs. Find the probability that a random sample of 16 bulbs will have an average life of less than 775 hrs.

Solution

Let X_i denotes the life of the light bulb, then we have

Mean $\mu = 800,\ \sigma = 40$

Let us assume that $\underline{x}$ denote the mean lifetime of 16 bulbs.

By central limit theorem, we have $\underline{x}$ follows a normal distribution with mean μ and variance $\frac{\sigma^2}{n}$.

i.e. $\underline{x}$ follows $N\left(\mu, \frac{\sigma}{\sqrt{n}}\right) \Rightarrow$ $\underline{x}$ follows $N\left(800, \frac{40}{\sqrt{16}}\right)$

We have to find $P(\underline{x} < 775)$

Let $z = \frac{\underline{x}-\mu}{\sigma/\sqrt{n}}$ is a standard normal variable.

$$z = \frac{\underline{x} - 800}{40/\sqrt{16}}$$

Now, $P(\underline{x} < 775) = P\left(\frac{\underline{x}-800}{40/\sqrt{16}} < \frac{775-800}{40/\sqrt{16}}\right)$

$$= P\left(z < -\frac{25 \times 4}{40}\right)$$

$$= P(z < -2.5)$$

$$= 0.5 - 0.4938$$

$$= 0.0062$$

The probability that a random sample of 16 bulbs will have an average life of less than 775 hours is 0.0062 (or 0.62%).

Illustration.4.3.3

The life time of a certain brand of a Tube light may be considered as a random variable with mean 1200 hr and standard deviation 250 hrs. Find the probability, using central limit theorem, that the average life time of 60 lights exceeds 1250 hrs.

Solution

Let X_i denotes the life time of the light, then we have

Mean $\mu = 1200,\ \ \sigma = 250$

Let us assume that $\underline{x}$ denote the mean life time of 60 lights.

By central limit theorem, we have $\underline{x}$ follows a normal distribution with mean μ and variance $\frac{\sigma^2}{n}$.

i.e. $\underline{x}$ follows $N\left(\mu, \frac{\sigma}{\sqrt{n}}\right) \Rightarrow$ $\underline{x}$ follows $N\left(1200, \frac{250}{\sqrt{60}}\right)$

We have to find $P(\underline{x} > 1250)$

Let $z = \frac{\underline{x}-\mu}{\sigma/\sqrt{n}}$ is a standard normal variable.

$$z = \frac{\underline{x} - 1200}{250/\sqrt{60}}$$

Now, $P(\underline{x} > 1250) = P\left(\frac{\underline{x}-1200}{\frac{250}{\sqrt{60}}} > \frac{1250-1200}{\frac{250}{\sqrt{60}}}\right)$

$$= P\left(z > \frac{50\sqrt{60}}{250}\right)$$

$$= P(z > 1.55)$$

$$= P(0 < z > 3) - P(0 < z < 1.55)$$

$$= 0.5 - (Area\ from\ 0\ to\ 1.55)$$

$$= 0.5 - 0.4394$$

$$= 0.0606$$

The probability that the average lifetime of 60 tube lights exceeds 1250 hours is 0.0606 (or 6.06%).

Illustration.4.3.4

The length of time, in hours, it takes an 'over 40' group of people to play one soccer match is normally distributed with a mean of 2 hours and a standard deviation of 0.5 hours. A sample of size $n = 50$ is drawn randomly from the population. Find the probability that the sample mean is between 1.8 hours and 2.3 hours.

Mean $\mu = 2,\ \sigma = 0.5$

Let us assume that $\underline{x}$ denote the sample mean of 50 people.

By central limit theorem, we have $\underline{x}$ follows a normal distribution with mean μ and variance $\frac{\sigma^2}{n}$.

i.e. $\underline{x}$ follows $N\left(\mu, \frac{\sigma}{\sqrt{n}}\right) \Rightarrow$ $\underline{x}$ follows $N\left(2, \frac{0.5}{\sqrt{50}}\right)$

We have to find $P(1.8 < \underline{x} < 2.3)$

Let $z = \frac{\underline{x}-\mu}{\sigma/\sqrt{n}}$ is a standard normal variable.

$$z = \frac{\underline{x} - 2}{\frac{0.5}{\sqrt{50}}}$$

Now, $P(1.8 < \underline{x} < 2.3) = P\left(\frac{1.8-2}{\frac{0.5}{\sqrt{50}}} < \frac{\underline{x}-2}{\frac{0.5}{\sqrt{50}}} > \frac{2.3-2}{\frac{0.5}{\sqrt{50}}}\right)$

$$= P(-2.8284 < z > 4.2426)$$

$$= 0.4976 + 0.4999$$

$$= 0.9975$$

The probability that the sample mean is between 1.8 hours and 2.3 hours is 0.9977 (or 99.77%).

Illustration.4.3.5

A computer generates 100 random numbers which are uniformly distributed between 0 and 1. Find the probability that their sum is at least 50.

Solution

For a uniformly distribution Mean $= \frac{1}{1-0} = 1$, variance $= \frac{(1-0)^2}{12} = \frac{1}{12}$

We have to find $P(S_n > 50)$

$$z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}}$$

$$= \frac{S_n - 100 \times \frac{1}{2}}{\frac{1}{12}\sqrt{100}}$$

$$P(S_n > 50) = P\left(\frac{S_n - 100 \times \frac{1}{2}}{\frac{1}{12}\sqrt{100}} > \frac{50 - 100 \times \frac{1}{2}}{\frac{1}{12}\sqrt{100}}\right)$$

$$= P\left(Z_n > \frac{50 - 100 \times \frac{1}{2}}{\frac{1}{12}\sqrt{100}}\right)$$

$$= P(Z_n > 0)$$

$$= 0.5$$

The probability that the sum of 100 uniform random numbers is at least 50 is 0.5 (or 50%).

Illustration.4.3.6

Let $X_1, X_2, \ldots \ldots X_n$ be an independent Poisson random variable with mean 1. Use central limit theorem to approximate $P(\sum_{i=1}^{20} X_i > 15)$.

Solution

Given Information:

We have independent Poisson(1) random variables:

$$X1, X2, \cdots, X120 \sim Poisson(\lambda = 1)$$

- We define the sum: S = $\sum_{i=1}^{120}$ X_i
- Since the sum of independent Poisson random variables follows a Poisson distribution, we get: $S \sim Poisson(\lambda = 120 \times 1 = 120)$
- We need to approximate: P(S>15)

Apply the Central Limit Theorem (CLT)

For a **Poisson(λ)** distribution:

- **Mean**:
  E[S] = λ = 120
- **Variance**:
  Var(S) = λ = 120
- **Standard deviation**:
  $\sigma_S = 120 \approx 10.95$

By **CLT**, SSS is approximately normally distributed:

$$S \sim N(120, 10.95^2)$$

Now, we standardize S=15:

$$Z = \frac{15-120}{10.95} = \frac{-105}{10.95} \approx -9.59$$

Find the Probability from the Z-Table

From standard normal tables:

$$P(Z<-9.59) \approx 0$$

Thus,

$$P(S>15) = 1-P(S\leq 15) \approx 1-0 = 1$$

Recap

- The central limit states that the distribution of sample means approaches a normal distribution as the sample size increases.
- Continuous probability function

Objective Questions

1. What does the Central Limit Theorem state about the distribution of sample means?
2. Which condition is necessary for the Central Limit Theorem to hold?
3. As the sample size increases, what happens to the standard deviation of the sample mean?
4. According to the Central Limit Theorem, what happens when sample size increases?
5. Which distributions can be approximated using the Central Limit Theorem when sample size is large?
6. If the population has a mean μ and a standard deviation σ, what will be the standard deviation of the sample mean ($\bar{X}$\) for a sample of size n?
7. The Central Limit Theorem is important because it allows us to:
8. What is the minimum sample size often recommended for the Central Limit Theorem to be applicable?
9. As per the Central Limit Theorem, the mean of the sampling distribution of the sample mean is equal to:

Answer

1. Approaches a normal distribution as sample size increases.
2. The sample size should be sufficiently large.
3. Decreases
4. The sample mean distribution becomes approximately normal
5. Any probability distribution, regardless of its shape
6. $\frac{\sigma}{\sqrt{n}}$
7. Use normal probability models to make inferences about population parameters
8. 30
9. The population mean μ

UNIT- 4

ESTIMATION AND CONFIDENCE LEVEL

Learning outcome

At the end of this unit, the student will be able to:

- Define estimation and confidence level in the context of statistical inference.
- State the difference between point estimation and interval estimation.
- Identify the key components of a confidence interval, including sample statistic, margin of error, and confidence level.
- Describe how increasing or decreasing the confidence level affects the width of the confidence interval.

Pre-requisites

Estimation and confidence level are foundational concepts in statistical analysis, requiring a solid understanding of probability, sampling methods, and data variability. A thorough grasp of probability helps in interpreting the likelihood of events and forms the basis for determining confidence intervals. Familiarity with sampling methods, such as random sampling or stratified sampling, is essential for ensuring that the data collected is representative of the larger population, thereby increasing the accuracy of estimates. Understanding data variability, including concepts like standard deviation and variance, is crucial for assessing the precision of estimates and determining appropriate confidence levels. Estimation and confidence level are foundational concepts in statistical analysis, requiring a solid understanding of probability, sampling methods, and data variability. A thorough grasp of probability helps in interpreting the likelihood of events and forms the basis for determining confidence intervals. Familiarity with sampling methods, such as random sampling or stratified sampling, is essential for ensuring that the data collected is representative of the larger population, thereby increasing the accuracy of estimates. Understanding data variability, including concepts like standard deviation and variance, is crucial for assessing the precision of estimates and determining appropriate confidence levels. Additionally, knowledge of statistical distributions, such as the normal or t-distribution, is vital for constructing confidence intervals and interpreting the reliability of estimations. These concepts collectively enable the effective application of

estimation techniques and the accurate interpretation of confidence levels in various analytical contexts.

Key Concepts

Point estimation, Interval Estimation, Confidence interval

Discussion

In the estimation of parameters, the objective is to obtain a precise value for the parameter itself. The process of estimation comes in two forms: point estimation and confidence interval. The estimators must possess fundamental properties such as unbiasedness, consistency, efficiency, and sufficiency which characterize the quality of estimators.

In statistical analysis, estimating population parameters is pivotal, particularly when direct observation of certain numerical characteristics is not possible. Instead, through the observation of related random variables, we seek to develop methods that utilize sample data to gain insights into these unobservable characteristics.

Starting with a random sample of size "n" from a population, estimation theory assumes the observations to be random, allowing us to identify the probability distribution dependent on parameters of interest for further analysis. Understanding and determining these parameters are essential as they define the distribution and guide the estimation process.

Suppose we are to find the average life of some electric switches in a large shipment of switches. Testing every switch to failure will give an exact average life, but the knowledge will most certainly be useless since we won't have any switches left. The correct procedure is to take a sample of n switches, test them to failure and take their average life $\underline{x}$. Then, we can take it as an estimate of the population mean.

All terms in any field of statistical investigation constitute a population. A population consisting of a finite number of items is called a finite population. A population consisting of a very large number of items or an infinite number of items is called an infinite population.

A complete enumeration of all items of a population is called census. But when the population is very large or infinite, census become difficult because it involves a great deal of time, money and energy. Therefore, for the purpose of statistical studies one can select a few items of this population according to some rules for studying the characteristics of the population. These selected items constitute a sample. Sampling may be defined as the process of obtaining information about the population by examining a part of it. One of the main objectives of statistical analysis is to draw information about some characteristics or attributes of the population from the corresponding attributes of the sample. Therefore, the sample should be reliable so that it may result in valid conclusions.

Two most important types of problems of inference in statistics are

i) Estimation ii) Testing of Hypothesis

4.4.1 Estimation

In the estimation of parameters, the objective is to obtain a precise value for the parameter itself. The process of estimation comes in two forms: point estimation and confidence interval. The estimators must possess fundamental properties such as unbiasedness, consistency, efficiency, and sufficiency which characterize the quality of estimators.

4.4.2 Point estimation

Suppose a retail company is considering setting the price for a new product they are about to launch. The population in this scenario would encompass all similar products in the market, both from competitors and those with similar features. The marketing team of the company decides to collect a sample of recent pricing data for similar products from five competitors in the market. They then adjust these prices based on features, brand reputation, and other factors to align them closely with the new product they plan to introduce.

Subsequently, the marketing team calculates the mean of the adjusted prices in their sample and recommends a pricing strategy of ₹99 for the new product. By doing so, they are making an inference about the mean price of the population. To assess the reliability of this inference, the marketing team should consider various factors: the relatively small sample size, potential fluctuations in market pricing over time, or the possibility that the new product might have unique features that set it apart from the sampled products. This thorough evaluation will help ensure that the pricing decision is based on a well-founded and informed assessment of the population's characteristics. An estimate of a population parameter given by a single number is called a point estimate of the parameter.

4.4.3 Interval Estimation

In the context of the earlier example, interval estimation refers to the process of determining a range of values within which a population parameter is likely to fall. In this case, the company wants to estimate the true mean price of similar products in the market to help set the price for their new product.

Instead of providing a single point estimate, such as suggesting that the mean price is exactly ₹99, interval estimation offers a range of values that is expected to contain the true population mean with a certain level of confidence. For instance, the marketing team might use statistical methods to calculate a confidence interval, which is a range of prices within which they believe the true mean price of similar products lies.

Let us say the marketing team calculates a 95% confidence interval for the mean price to be ₹95 to ₹103 based on their sample data and adjustments. This means that they are 95% confident that the true mean price of similar products in the market falls within this interval.

An estimate of a population parameter given by two numbers between which the parameter may be considered to lie is called an interval estimate of the parameter. Interval estimation acknowledges the inherent uncertainty in making population inferences from a sample. By providing a range of values instead of a single point estimate, the marketing team of the company gives a more comprehensive picture of the potential variability in the parameter being estimated. This approach helps decision-makers make informed choices and manage risks associated with their decisions, considering the potential fluctuations in the population characteristics.

4.4.6 Properties of estimators

A good estimator should possess the following properties

- **Unbiasedness:** An estimator is unbiased if, on average, it yields an estimate that is equal to the true value of the parameter being estimated. In other words, the expected value of the estimator matches the population parameter it aims to estimate. An unbiased estimator minimizes the systematic errors that may arise during the estimation process.
- **Consistency:** An estimator is consistent if it converges in probability to the true parameter value as the sample size increases. In simpler terms, as more data are collected, the estimator becomes more accurate and approaches the true population parameter. Consistency ensures that the estimator becomes reliable when working with larger samples.
- **Efficiency:** An estimator is efficient if it has a smaller variability (lower variance) than other estimators for the same parameter. An efficient estimator strikes a balance between providing accurate estimates and having minimal variability. In essence, it provides the most information about the parameter using the available data.
- **Sufficiency:** An estimator is sufficient if it captures all relevant information about the parameter contained in the sample data. A sufficient estimator reduces the entire dataset down to a manageable summary, allowing for efficient estimation without unnecessary redundancy. This property is particularly important in cases where data are complex and extensive.

4.4.7 Confidence Level

The probability of an estimator to lie between the intervals is called the confidence level.

A 95% confidence level of a parameter θ to lie in the interval (c_1, c_2). We mean

$P[c_1 < \theta < c_2] = 0.95$. We usually denote the confidence level by $1 - \alpha$ and α is called level of significance (LOS). That is 95% confidence level means 5% level of significance. Thus, in an interval estimates, we determine two constants c_1 and c_2 such that $P[c_1 < \theta < c_2] = 1 - \alpha$ where α is the significance level. The interval (c_1, c_2) is called confidence interval and the limits c_1 and c_2 are called confidence limits.

4.4.7.1 Confidence Intervals for Mean

From central limit theorem, we know that $Z = \frac{\bar{X}-\mu}{SE}$ approaches standard normal distribution as $n > 30$.

Suppose we want to find the interval estimation of μ with confidence level 95%.

i.e. $1 - \alpha = 0.95 \Rightarrow \alpha = 0.05$

Thus we want to find c_1 and c_2 such that $P[c_1 < Z < c_2] = 0.95$.

But from the standard normal table

$$P[-1.96 < Z < 1.96] = 0.95.$$

$$P[-1.96 < \frac{\bar{X}-\mu}{SE} < 1.96] = 0.95.$$

If the statistic S is the sample mean $\underline{X}$, then 95% and 99% confidence limits for estimation of the population mean μ are given by $\underline{X} \pm 1.96\, SE$, $\underline{X} \pm 2.58\, SE$, respectively.

If the confidence limit are not specified, we take the confidence limit as

$\underline{X} \pm Z_{\frac{\alpha}{2}} SE$ where $Z_{\frac{\alpha}{2}}$ is the critical value or significant value.

Confidence Level	90%	95%	99%
$Z_{\frac{\alpha}{2}}$	1.645	1.96	2.58

4.4.7.2 Confidence Interval for Small Samples

It is to be noted that when the sample is small (i.e.,n≤30), the t test has to be conducted.

If the population standard deviation is not known and the sample is small, we have the confidence limit $\underline{X} \pm t_{\frac{\alpha}{2}} SE$ where $t_{\frac{\alpha}{2}}$ is that value from the table of t-distribution for which $P[-t_{\frac{\alpha}{2}} < t < t_{\frac{\alpha}{2}}] = 1 - \alpha$ with $n - 1$ degree of freedom.

4.4.7.3 Confidence Interval for Proportions

Suppose that the statistic S is the proportion of "successes" in a sample of size n ≧ 30 drawn from a binomial population in which p is the proportion of successes (i.e. the probability of success).

Then the confidence limits for p are given by $p \pm Z_{\frac{\alpha}{2}} SE$ where P denotes the proportion of successes in the sample of size n and the SE= $\sqrt{\frac{pq}{n}}$ where q denotes the proportion of failure in the sample

Illustration.4.4.1

A sample of 50 items taken from a population with standard deviation 16 gave a mean 52.5. Find a 95% confidence interval of the population mean?

Solution

Here $n = 50,\ \sigma = 16,\ \underline{x}$=52.5

Since n>30, the sample is large. Hence SE= $\frac{\sigma}{\sqrt{n}} = \frac{16}{\sqrt{50}} = 2.26$

For 95% confidence interval is

$$\left(\underline{X} - \frac{1.96\sigma}{\sqrt{n}}, \underline{X} + 1.96\frac{\sigma}{\sqrt{n}}\right)$$

i.e. $(52.5 - 1.96 \times 2.26,\ 52.5 + 1.96 \times 2.26)$

i.e.$(48.07,\ 56.93)$

Illustration.4.4.2

A sample of 100 items taken from a population with standard deviation 1.2kg. gave a mean 7.4 kg. Find a 95% confidence interval of the population mean?

Solution

Here $n = 100,\ \ s = 1.2,\ \ \underline{x} = 7.4$.

Since n>30, the sample is large. Hence SE= $\frac{s}{\sqrt{n}} = \frac{1.2}{\sqrt{100}} = 0.12$

For 95% confidence interval is

$$\left(\underline{X} - \frac{1.96s}{\sqrt{n}}, \underline{X} + 1.96\frac{s}{\sqrt{n}}\right)$$

i.e. $(7.4 - 1.96 \times 0.12,\ 7.4 + 1.96 \times 0.12)$

i.e.$(7.1648,\ \ 7.6352)$

Illustration.4.4.3

The mean operating life for a random sample of 15 bulbs taken from a population with standard deviation 500hrs is 8900 hrs. Find a) 95% confidence limit b) 90% confidence limit for the population mean.

Solution

n = 15, $\underline{X} = 8900$, σ = 500,

Since $n < 30$, the sample is small and the population standard deviation is known, we use normal distribution.

Confidence $limit$ $\ \ \underline{X} \pm Z_{\frac{\alpha}{2}} SE$

i) $Z_{\frac{\alpha}{2}} = 1.96$

$$SE = \frac{\sigma}{\sqrt{n}} = \frac{500}{\sqrt{15}} = 129.1$$

Confidence Limit

$$\left(\underline{X} - \frac{1.96\sigma}{\sqrt{n}}, \underline{X} + 1.96\frac{\sigma}{\sqrt{n}}\right)$$

$$(8900 - 1.96 \times 129.1\ , 8900 + 1.96 \times 129.1)$$

$$(8647, 9153)$$

i) $Z_{\frac{\alpha}{2}} = 1.645$

$$SE = \frac{\sigma}{\sqrt{n}} = \frac{500}{\sqrt{15}} = 129.1$$

Confidence Limit

$$\left(\underline{X} - \frac{1.645\sigma}{\sqrt{n}}, \underline{X} + 1.645\frac{\sigma}{\sqrt{n}}\right)$$

$$(8900 - 1.645 \times 129.1\ , 8900 + 1.645 \times 129.1)$$

$$(8687, 9113)$$

Illustration.4.4.4

The mean operating life for a random sample of 10 light bulbs is 4000 hours with a standard deviation 200hrs. Find a) 95% confidence interval for the population mean.

Solution

n = 10, $\underline{X} = 4000$, σ = 200,

Since $n < 30$, the sample is small. σ is not known we use t- distribution with

$n - 1 = 10 - 1 = 9$ degree of freedom.

Confidence $limit$ $\quad \underline{X} \pm t_{\frac{\alpha}{2}} SE$

$$SE = \frac{s}{\sqrt{n}} = \frac{200}{\sqrt{10}} = 63.25$$

$t_{\frac{\alpha}{2}} = 2.262$ at 9 degree of freedom

Confidence Limit

$$\left(\underline{X} - t_{\frac{\alpha}{2}} SE, \underline{X} + t_{\frac{\alpha}{2}} SE\right)$$

$$(400 - 2.262 \times 63.25\ , 400 + 2.262 \times 63.25)$$

$$(3857, 4143)$$

Illustration 4.4.5

In a sample of 400 people, 172 were males. Estimate the population proportion at 95% confidence level.

Solution

n = 400, $p = \frac{172}{400} = 0.43 \ \ and\ q = \frac{328}{400} = 0.57,\ \ Z_{\frac{\alpha}{2}} = 1.96$

$$SE = \sqrt{\frac{pq}{n}} = \sqrt{\frac{0.43 \times 0.57}{400}} = 0.0248$$

Confidence Limit is $P \ \pm \ Z_{\frac{\alpha}{2}} SE$

$$\text{i.e. } 0.43 \pm 1.96 \times 0.0248$$

$$\text{i.e. } (0.3814,\ 0.4786)$$

Illustration 4.4.6

In a sample of 450 industrial accidents it was found that 230 were due to unsafe working condition. Construct 95% confidence interval for the corresponding true proportion.

Solution

$n = 450$, $p = \frac{230}{450} = 0.51$ and $q = \frac{220}{450} = 0.49$, $Z_{\frac{\alpha}{2}} = 1.96$

$$SE = \sqrt{\frac{pq}{n}} = \sqrt{\frac{0.51 \times 0.49}{450}} = 0.0236$$

Confidence Limit is $P \pm Z_{\frac{\alpha}{2}} SE$

$$\text{i.e. } 0.51 \pm 1.96 \times 0.0236$$

$$\text{i.e. } (0.4638,\ \ 0.5563)$$

Recap

- Estimation: In the estimation of parameters, the objective is to obtain a precise value for the parameter itself.
- Point Estimation: An estimate of a population parameter given by a single number is called a point estimate of the parameter.
- Estimation: In the estimation of parameters, the objective is to obtain a precise value for the parameter itself.
- Point Estimation: An estimate of a population parameter given by a single number is called a point estimate of the parameter.
- Interval estimate: An estimate of a population parameter given by two numbers between which the parameter may be considered to lie is called an interval estimate of the parameter.
- Properties of estimators
- Confidence Level: The probability of an estimator to lie between the intervals is called the confidence level

Objective Questions

1. What is the primary purpose of estimation in statistics?
2. Give an example of a point estimate?
3. What does a 95% confidence level mean?
4. Which distribution is commonly used for constructing confidence intervals for the population mean when the sample size is small (<30) and population variance is unknown?
5. If the confidence interval for a population mean is (50, 70), what can be concluded?
6. If a hypothesis test results in a p-value of 0.03 and the confidence level is 95%, what should be concluded?

Answer

1. To approximate a population parameter based on sample data
2. Sample mean used to estimate population mean
3. If we take many samples, 95% of the confidence intervals will contain the true population parameter
4. Student's t-distribution
5. We are confident that the population mean lies between 50 and 70 with a given probability
6. The null hypothesis should be rejected

Assignment

1. Explain the concepts of **point estimation** and **interval estimation** with examples.
2. Compare and contrast the **sample mean** and **population mean** in estimation.
3. What is the significance of the **confidence level** in statistical estimation?
4. Explain the role of the **margin of error** in confidence intervals.
5. Why is a **larger sample size** preferred in estimation?
6. A survey reports a **95% confidence level** for an estimated proportion. Explain what this means in layman's terms.
7. How would the confidence interval change if we increase the confidence level from **95% to 99%**?
8. A random sample of **50 students** has an average test score of **75** with a standard deviation of **10**. Compute a **95% confidence interval** for the true mean test score of all students.
9. In a poll of **500 people**, **60%** support a new policy. Construct a **95% confidence interval** for the true proportion of supporters.

10. A company surveys **200 customers** and finds that the average satisfaction rating is **4.2 out of 5**, with a standard deviation of **0.5**. Compute a **95% confidence interval** and explain its significance in decision-making.
11. A factory tests **samples of light bulbs** and finds an average lifespan of **1,000 hours** with a standard deviation of **50 hours**. Compute a **90% confidence interval** to estimate the true average lifespan of all light bulbs produced.
12. Collect real-world data (e.g., **average commute times, monthly expenses, or student GPAs**) and compute a **confidence interval** for the population mean. Interpret your findings and discuss how the confidence level affects the results.

Suggested reading

1. John Wiley & Sons, Inc., NY, USA, 2010 3. Montgomery, D. C. and G. C. Runger. Applied Statistics and Probability for Engineers. 5th Edition. John Wiley & Sons, Inc., NY, USA, 2011.
2. Kolman, Busby, Ross and Rehman, Discrete Mathematical Structures for Computer Science, Pearson Education,6th Edition, 2017.
3. Erwin Kreyszig, Advanced Engineering Mathematics, Wiley India, 9th Edition, 2011.

References

1. K Huffman and R Kunze, Linear Algebra, Pearson Education, 2nd Edition, 2005.
2. Thomas M. Apostol, Calculus, Wiley, 2nd Edition, 1991 ISBN 960-07-0067-2.
3. Michael Spivak. Calculus, publish or Perish, 2008, ISBN 978-0914098911.
4. Ross L. Finney, Maurice D.Weir. and Frank R. Giordano. Thomas's Calculus, Pearson l2th Edition 2009.
5. Bendat, J. S. and A. G. Piersol. Random Data: Analysis and Measurement Procedures. 4th Edition.

Hypothesis Testing

UNIT 1
Null and Alternative Hypothesis

Learning outcome

After completing this unit, the learner will be able to:

- Recall the definition of the null hypothesis and alternative hypothesis.
- Recognize examples of null and alternative hypotheses.
- State the purpose of the null and alternative hypotheses in statistical testing.
- Recall the definition of a simple and composite hypothesis.
- Label the null and alternative hypotheses for simple statistical problems.

Pre-requisites

Before understanding null and alternative hypotheses, learners should have a basic knowledge of statistical concepts, including population and sample, variables, and types of data. Familiarity with descriptive statistics like mean, median, and standard deviation and the concept of probability is essential, as hypothesis testing relies on probability to draw conclusions. Additionally, an understanding of sampling methods and the idea of sampling distribution helps in comprehending how sample data can be used to make inferences about a larger population.

Key Concepts

Null hypothesis, Alternative hypothesis

Discussion

The goal of hypothesis testing is to determine if there is enough evidence to support the initial assumption or claim about a population parameter. This involves using statistical analysis to see if the collected data aligns with the proposed hypothesis. Based on the results, a decision is made either to accept the hypothesis or reject it in favour of an alternative explanation. This process helps guide decision-making by relying on data rather than assumptions.

Grasping the Concept of a Hypothesis

A hypothesis is a well-thought-out guess about a parameter formed based on prior knowledge or observation. It acts as a starting point for statistical investigation. To test a hypothesis, relevant data is collected and analyzed. The results from the data analysis help us decide if we can support or reject the original assumption. This step-by-step process is the foundation of hypothesis testing.

Scientific Nature of a Hypothesis

In statistical terms, a hypothesis is defined as a statement that can be tested using scientific methods. For example, a hypothesis might claim that "students who study for more than two hours daily score higher than those who don't." By gathering and analyzing relevant data, this statement can be tested to see if it holds true or not. Testing ensures that conclusions are based on objective evidence rather than subjective opinion.

For example, 'Those who smoke have a greater risk of having cancer'. A statistical hypothesis is a statement about a population that we want to test based on the result obtained from a random sample. That is, it is a statement about a parameter that we test using the value of a statistic. For this purpose, the hypothesis should possess the following characteristics.

i) It should be clear and precise

ii) It should be capable of being tested

iii) It should state the relationship between variables.

5.1.1. Simple and Composite Hypothesis

Hypotheses can be categorized into two main types: simple and composite. A simple hypothesis confines itself to a singular distribution for X. In this scenario, the hypothesis defines one specific outcome or pattern that X's distribution should adhere to.

On the other hand, a composite hypothesis entertains the possibility of multiple distributions for X. This means that the hypothesis encompasses a range of potential distributional patterns for X. The composite hypothesis acknowledges the variability in X's behaviour and contemplates various alternatives within its scope.

5.1.2. Null and Alternative Hypothesis

When we attempt to make a decision about the population on the basis of sample information, we have to make assumptions or guesses about the nature of the population involved or about the value of some parameter of the population. Such assumptions, which may or may not be true, are called statistical hypotheses. We set up a hypothesis that assumes that there is no significant difference between the sample statistic and the corresponding population parameter or between two sample statistics. Such a hypothesis of no difference is called null hypothesis and is represented as H_0. A hypothesis that is complement to the null hypothesis is the alternative hypothesis and is denoted by H_1. A procedure for deciding whether to accept or to reject a null hypothesis or hence to reject or accept the alternative hypothesis, is called test of hypothesis.

Given the contradictory nature of the null and alternative hypotheses, the examination of evidence becomes crucial in determining whether there is sufficient reason to reject the null hypothesis. The evidence in question is derived from the sample data.

For example, suppose one wishes to test whether the mean of a population μ is 100. We can take $H_0: \mu = 100$. If this null hypothesis is rejected, then it would mean any of the three possible alternatives.

i.e. $\mu \neq 100, \quad \mu < 100, \quad \mu > 100$

Thus we can test $H_0: \mu = 100$ against

i) $H_0: \mu = 100, \ H_1: \mu \neq 100$

ii) $H_0: \mu = 100, \ H_1: \mu < 100$

iii) $H_0: \mu = 100, \ H_1: \mu > 100$

The hypothesis $\mu = 100$ is called simple hypothesis and the hypothesis , $\mu < 100, \quad \mu > 100$ are called composite hypothesis.

Recap

- Hypothesis - statement which is capable of being tested by scientific methods.
- Simple hypothesis - the hypothesis defines one specific outcome.
- Composite hypothesis - the hypothesis defines multiple distributions for *X*.
- Null hypothesis - there is no significant difference between the sample statistic and the corresponding population parameter or between two sample statistics.
- Alternative hypothesis -A hypothesis that is complement to the null hypothesis.

Objective Type Questions

1. What is the goal of hypothesis testing in statistical analysis?
2. Define a hypothesis in simple terms.
3. What are the steps involved in testing a hypothesis?
4. How does a hypothesis help in decision-making?
5. What are the characteristics a hypothesis should possess?
6. What is meant by a simple hypothesis?
7. What is a composite hypothesis?
8. Define a null hypothesis and its notation.
9. What is an alternative hypothesis and its notation?
10. How is evidence used to decide whether to accept or reject a null hypothesis?

Answer to Objective Type Questions

1. To determine if there is enough evidence to support or reject an assumption about a population parameter.
2. A well-thought-out guess or assumption about a parameter based on prior knowledge or observation.
3. Form hypothesis, collect data, analyze data, make a decision.
4. Clear and precise, capable of being tested, states relationship between variables.
5. Simple hypothesis specifies a single distribution; composite hypothesis allows for multiple distributions.
6. Null hypothesis assumes no significant difference; alternative hypothesis is its complement.
7. H0 (null hypothesis), H1 (alternative hypothesis).
8. It represents the hypothesis of no significant difference.
9. To test claims about a population parameter.
10. Sample data.

Assignment Question

1. Explain the key differences between simple and composite hypotheses with examples.
2. Discuss the steps involved in hypothesis testing and explain how they guide decision-making.
3. Describe the characteristics a hypothesis should have to be scientifically testable.
4. Differentiate between the null hypothesis and the alternative hypothesis, providing suitable examples for each.
5. Using a real-world scenario, formulate a hypothesis and outline how you would test it using statistical methods.

Suggested reading

1. K Huffman and R Kunze, Linear Algebra, Pearson Education, 2nd Edition, 2005.
2. Thomas M. Apostol, Calculus, Wiley, 2nd Edition, 1991 ISBN 960-07-0067-2.
3. Michael Spivak. Calculus, publish or Perish, 2008, ISBN 978-0914098911.
4. Ross L. Finney, Maurice D.Weir. and Frank R. Giordano. Thomas's Calculus, Pearson l2th Edition 2009.
5. Bendat, J. S. and A. G. Piersol. Random Data: Analysis and Measurement Procedures. 4th Edition.
6. John Wiley & Sons, Inc., NY, USA, 2010 3. Montgomery, D. C. and G. C. Runger. Applied Statistics and Probability for Engineers. 5th Edition. John Wiley & Sons, Inc., NY, USA, 2011.

7. Kolman, Busby, Ross and Rehman, Discrete Mathematical Structures for Computer Science, Pearson Education,6th Edition, 2017.
8. Erwin Kreyszig, Advanced Engineering Mathematics, Wiley India, 9th Edition, 2011.

UNIT 2
Type I and Type II error

Learning outcome

After completing this unit, the learner will be able to:

- Define Type I and Type II errors in hypothesis testing.
- Explain the concepts of critical region and level of significance (LOS).
- Describe how the level of significance impacts hypothesis testing.
- Understand the power of a test and its importance in decision-making.
- Differentiate between one-tailed and two-tailed tests.
- Interpret p-values and their role in hypothesis testing decisions.

Pre-requisites

Before understanding Type I and Type II errors, it is essential to have a foundational knowledge of hypothesis testing and statistical decision-making. Learners should be familiar with the null hypothesis and alternative hypothesis, as these errors occur in the process of testing them. A basic understanding of significance levels and p-values is crucial, as Type I error is directly related to α (the probability of rejecting a true null hypothesis). Concepts like sample size, confidence intervals, and probability distributions also play a role in determining the likelihood of these errors.

Keywords:

Type I error, Type II error

Discussion

Hypothesis testing should be conducted with utmost care. This is due to its implications on decision-making and the potential consequences of drawing incorrect conclusions. Errors in hypothesis testing can lead to wrong decisions that impact real-world scenarios. In testing a

hypothesis, we can make two types of mistakes. There are two possible types of errors that one can made in testing a hypothesis.

5.2.1. Type I error

A **Type I error** is rejecting a true hypothesis. i.e. Rejecting the null hypothesis H_0 when it is true. The probability of type I error is denoted by α and is called the level of significance.

For example, imagine a new COVID-19 test that is designed to detect whether a person is infected. The null hypothesis (H_0) states that the person does not have COVID-19, while the alternative hypothesis (H_1) states that the person does have COVID-19.

If the test incorrectly detects COVID-19 in a healthy person (when they are actually not infected), this is a Type I error—a false positive result. As a result, the person may be unnecessarily isolated, given unneeded medication, or experience anxiety, even though they were never sick.

5.2.2. Type II error

A **Type II error** is not rejecting a false hypothesis. Accepting the null hypothesis H_0 when it is false. The probability of type II error is denoted by β and is called the power of the test.

5.2.3. Critical region

The **critical region** of a hypothesis test encompasses the values of the test statistic that would lead to the rejection of the null hypothesis. This region is also referred to as the rejection region because it defines the range of extreme or significant values that signal a departure from the null hypothesis.

5.2.4. Level of Significance (LOS)

When the significance level is set, it corresponds to the probability of committing a Type I error (rejecting a true null hypothesis). This level is typically chosen based on the desired level of confidence in the test results.

The critical region's extent is directly linked to the significance level α. For example, if the significance level is set at 0.05, the critical region should account for 5% of the probability distribution associated with the reference distribution (often the Standard Normal Distribution).

Tests of significance are based on the fact that each null hypothesis can be tested with a particular type of statistical test. Every calculated statistic has a special distribution associated with it. The calculated value is then compared to the distribution to conclude whether the sample characteristics are different from what you would expect by chance.

Confidence intervals represent the best estimate of the range of the population value (or population parameter) based on the sample value (or sample statistic). A higher confidence interval (for example, a 99% confidence interval as compared with a 95% confidence interval) represents a greater degree of confidence, meaning that a wider range of values will be incorporated into the confidence interval.

5.2.5. Power of a test

In cases where the sample size is small, even if there exists a noticeable difference between sample means, it might not be statistically significant. In such situations, making definitive

conclusions about the population means becomes challenging. The lack of significance does not provide evidence in favor of the null hypothesis that the population means are equal, nor does it support the notion that the null hypothesis is false. Consequently, when an observed effect fails to attain significance, the outcome is inconclusive. In scenarios where resources are limited, you might prefer allocating funds to projects with a higher likelihood of yielding robust conclusions, allowing for more confident decision-making.

The **power of a test** is a crucial concept in statistical hypothesis testing, representing the probability of correctly identifying and rejecting a false null hypothesis.

5.2.6. One tail and two-tail tests

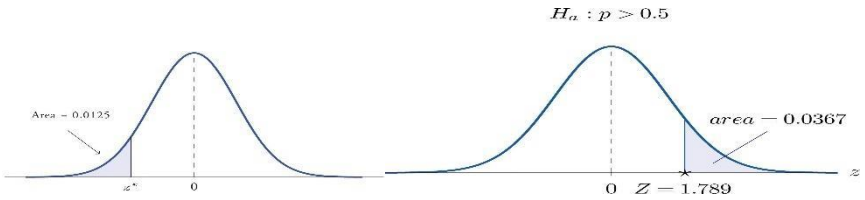
A test of any statistical hypothesis where the alternative hypothesis is one tailed (right tailed or left tailed) is called one tailed test.

For example, in a test for testing the mean of a population in a single tailed test we assume that the null hypothesis $H_0: \mu = \mu_0$ against the alternate hypothesis

$$H_1: \mu > \mu_0 \text{ (Right tailed)}$$

$$H_1: \mu < \mu_0 \text{ (Left tailed)}$$

is called one tailed test.



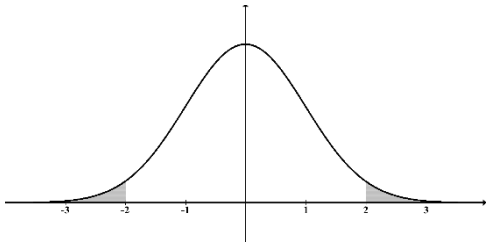
Left tailed curve Right tailed curve

In a test of any statistical hypothesis where the alternative hypothesis is two tailed, we assume that the null hypothesis

$$H_0: \mu = \mu_0$$

against the alternate hypothesis

$$H_1: \mu \neq \mu_0$$



Two tailed curve

5.2.6. p value

To take the decision about the null hypothesis on the basis of p-value, the p value is compared with given level of significance (α) and if p-value is less than or equal to ⟨ then we reject the null hypothesis and if the p-value is greater than ⟨ we do not reject the null hypothesis.

Since test statistic Z follows approximately normal distribution with mean 0 and variance unity, i.e. standard normal distribution and we also know that standard normal distribution is symmetrical about Z = 0 line therefore, if z represents the calculated value of Z then p-value can be calculated as follows:

For one-tailed test:

For $H_1: \theta > \theta_0$ (right-tailed test), p-value $= P[Z\ z]$

For $H_1: \theta < \theta_0$ (left-tailed test), p-value $= P[Z\ z]$

For two-tailed test: For H_1: $_0$ p-value $= 2P[Z \geq |z|$

For example, if test is right-tailed and calculated value of test statistic Z is 1.23 at 5% LOS

then p-value $= P[Z\ z] = P[Z\ 1.23] = 0.5 - P[0 < z < 1.23]$

$= 0.5 - 0.3907$ from Normal distribution table

$= 0.1093 > 0.05$

Thus, we do not reject the null hypothesis.

Recap

- Type I error: reject the null hypotheses when true.
- Type II error :Accepting the null hypothesis H_0 when it is false
- Critical region: Also called rejection region, the values of the test statistic that would lead to the rejection of the null hypothesis.
- Level of Significance : probability of type I error
- Power of a test: probability of type IL error
- p value: that describes how likely you are to have found a particular set of observations if the null hypothesis were true

Objective Type Questions

1. What is a Type I error?
2. What is a Type II error?
3. What we call the probability of making a Type I error?
4. What we call the probability of making a Type II error?
5. The rejection region is also known as,
6. What is the indication of smaller p-value?
7. If the p-value is greater than the significance level (α), what should we do?
8. What does a p-value represent?

Answers

1. Rejecting the null hypothesis when it is true
2. Accepting the null hypothesis when it is false
3. Level of significance
4. β
5. Power of a test
6. Critical region
7. Stronger evidence against the null hypothesis
8. Fail to reject the null hypothesis
9. The probability of obtaining the observed data if the null hypothesis is true

Assignment Questions

1. How does increasing the level of significance (α) affect the probability of Type II error (β)?
2. How does sample size affect Type I and Type II errors? Explain with examples

Suggested reading

1. K Huffman and R Kunze, Linear Algebra, Pearson Education, 2nd Edition, 2005.
2. Thomas M. Apostol, Calculus, Wiley, 2nd Edition, 1991 ISBN 960-07-0067-2.

3. Michael Spivak. Calculus, publish or Perish, 2008, ISBN 978-0914098911.
4. Ross L. Finney, Maurice D.Weir. and Frank R. Giordano. Thomas's Calculus, Pearson l2th Edition 2009.
5. Bendat, J. S. and A. G. Piersol. Random Data: Analysis and Measurement Procedures. 4th Edition.
6. John Wiley & Sons, Inc., NY, USA, 2010 3. Montgomery, D. C. and G. C. Runger. Applied Statistics and Probability for Engineers. 5th Edition. John Wiley & Sons, Inc., NY, USA, 2011.
7. Kolman, Busby, Ross and Rehman, Discrete Mathematical Structures for Computer Science, Pearson Education,6th Edition, 2017.
8. Erwin Kreyszig, Advanced Engineering Mathematics, Wiley India, 9th Edition, 2011.

Unit 3
Test Procedure

Learning outcome

After completing this unit, the learner will be able to:

- Define *population*, *sample*, and *sample size*.
- Identify the steps involved in hypothesis testing.
- State the conditions for rejecting or accepting the null hypothesis.
- Recognize different types of hypothesis tests (one mean, two means, one proportion, two proportions).
- Calculate the test statistic and compare it with critical values to make a decision on the null hypothesis.

Pre-requisites

The prerequisites for a test procedure include having a clear understanding of the test goals, the required equipment, and a well-defined environment. It is important to ensure that all necessary resources, such as tools, software, and personnel, are available. The test procedure should outline the steps, conditions, and expected outcomes, ensuring consistency and reliability. Additionally, it's essential to confirm that any dependencies, such as data or systems, are set up and functioning before starting the test. Preparing these elements ensures smooth execution and accurate results.

Keywords:

Test of significance, population, sample, hypothesis

Discussion

In a statistical investigation the interest usually lies in the assessment of the general magnitude and the study of variations with respect to one or more characteristics relating to individuals belonging to a group. The group of individuals under study is called population. The population may be finite or infinite.

It is obvious that for any statistical investigation complete enumeration of the population is rather impractical. For example, if we want to have an idea of the average monthly income of people in Kerala, we will have to enumerate all the earning individuals in the state which is rather a difficult task. Hence, we take a sample from the population.

A finite subset of statistical individuals in a population is called Sample. The number of individuals in a sample is called the sample size.

5.3.1. Procedure for testing of hypothesis

Applying a statistical test to any null hypothesis follows eight general steps:

1. Provide a statement of the null hypothesis.

2. Set the level of risk associated with the null hypothesis (significance level).

3. Select the appropriate test statistic $z = \frac{t-\mu}{SE}$

4. Compute the test statistic value (also known as the obtained value).

5. Determine the value (the critical value) needed for rejection of the null hypothesis using the appropriate table of critical values for that particular statistic.

If $|z| < 1.96$, H_0 may be accepted at 5% level of significance

If $|z| > 1.96$, H_0 may be rejected at 5% level of significance

If $|z| < 2.58$, H_0 may be accepted at 1% level of significance

If $|z| > 2.58$, H_0 may be rejected at 1% level of significance

For a single tail test (Right tailed or Left tailed) we compare the computed value of $|z|$ with 1.645 (at 5% LOS) and 2.33 (at 1% LOS) and accept or reject H_0 accordingly.

6. Compare the obtained value with the critical value.

7. If the obtained value is more extreme than the critical value, the null hypothesis must be rejected.

8. If the obtained value does not exceed the critical value, the null hypothesis cannot be rejected.

We are going to deal with the following tests

i) Testing the hypothesis concerning one mean
ii) Testing the hypothesis difference of two means
iii) Testing the hypothesis concerning one proposition
iv) Testing the hypothesis concerning difference of two proposition

5.3.2. Test of significance with respect to Mean

We usually conduct the test of significance with respect to mean when the sample consists of 'n' independent observations drawn from a normal distribution. The n independent random variables, follow a normal distribution with an unknown population mean (μ), and a known variance (σ^2).

Symbolically,

$$X \sim N(\mu, \sigma)$$

μ is unknown; σ^2 is known

The following are the key steps to conduct the test of significance with respect to mean.

Null Hypothesis $\mu = \mu_0$ (a value)

$$\underline{X} \sim \left(\mu, \frac{\sigma^2}{n}\right)$$

$$Z = \frac{\underline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

If the level of significance $\alpha = 0.05$

$$|Z| = \frac{|\underline{X} - \mu_0|}{\frac{\sigma}{\sqrt{n}}}$$

If $|Z| > 1.96 \Rightarrow$ Reject the Null Hypothesis.
If $|Z| < 1.96 \Rightarrow$ Do not reject the Null Hypothesis.

If the level of significance $\alpha = 0.01$

$|Z| > 2.58 \Rightarrow$ Reject Null Hypothesis
$|Z| < 2.58 \Rightarrow$ Do not reject Null Hypothesis

Illustration.5.3.1

A motor car company claims that their car average is 35 miles per gallon of Petrol. A random sample of 50 cars was tested and found to given an average of 32 miles per gallon. With a standard deviation of 1.2 gallons, test the null hypothesis μ = 35 against $H_1: \mu \neq 35$**.**

Solution

Given μ = 35, $\sigma = 1.2,\ \underline{x} = 32,\ n = 50$

Let $H_0: \mu = 35$
$H_1: \mu \neq 35$

As the sample size n is large , the test statistic is

$$Z = \frac{\underline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

$$Z = \frac{32 - 35}{\frac{1.2}{\sqrt{50}}}$$

$$Z = 17.25 > 1.96$$

H_1 is two sided and hence 5% LOS, from the normal table, $Z_{\frac{\alpha}{2}} = 1.96$. Since $|Z| > 1.96$, H_0 is rejected against H_1

Illustration.5.3.2.

Suppose that a census of city dwellers reveals an average family size of 4.2 with a standard deviation of 0.5. A random sample of 100 city side families reveals a family size of 4.29. We wish to test whether the family size in the city side is the same as in the city. Also find p value.

Solution

$$H_0: \mu = 4.2$$

$$H_1: \mu \neq 4.2$$

$$|Z| = \frac{\left|\underline{X} - 4.2\right|}{\frac{0.5}{10}} = \frac{|4.29 - 4.2|}{0.05} = \frac{0.09}{0.05} = \frac{9}{5} = 1.8$$

$1.8 < 1.96 \Longrightarrow$ Do not reject H_0 at 5% LOS.

Conclusion: Family size in the city side is same as in the city.

$$H_0: \mu = 4.2$$

$$H_1: \mu \neq 4.2$$

$$|Z| = \frac{|4.29 - 4.2|}{\frac{0.5}{10}} = 1.8 > 1.65$$

p-value

Since the test is two tailed test $p - \text{value} = 2P[Z \geq |z|]$

$$= 2P[Z \geq 1.8] = 2\{0.5 - P[0 < Z < 1.8] = 2\{0.5 - 0.4641 = 2 \times 0.0359 = 0.0718 > 0.05$$

do not reject the null hypothesis at 5% LOS.

Illustration.5.3.3.

A random sample of boots owned by 40 soldiers in a desert region showed an average life of 1.08 years with a standard deviation of 0.5 years. Under standard conditions, the boots are known to have an average life of 1.28 years. Is there a reason to assert at a level of significance of 0.01 that use in deserts causes the mean life of such boots to decrease. Also find p-value.

Solution

$$H_0: \mu = 1.28$$

$$H_1: \mu < 1.28$$

$|Z| = \frac{|\underline{X} - 1.28|}{\frac{0.5}{\sqrt{40}}} = \frac{|1.08 - 1.28|}{\frac{0.5}{6.32}} = \frac{0.2}{0.08} = 2.5 > 1.65$ at 1% LOS

Reject H_0

p-value

Since the test is left tailed, $P[Z\ z] = P[Z \leq -2.5] = 0.5 - P[0 < z < 2.5]$

$$= 0.5 - 0.4938$$

$$= 0.0062 < 0.01$$

reject null hypothesis

Illustration.5.3.4

It is claimed that a random sample of 100 tyres with mean life of 15269 kms is drawn from a population of tyres which has a mean life of 15200 kms and standard deviation is 124.8 kms. Test the validity of the claim at 5% LOS.

Solution

$$H_0: \mu = 15200$$

$$H_1: \mu \neq 15200$$

$$|Z| = \frac{|15269 - 15200|}{\frac{1248}{\sqrt{100}}} = \frac{69}{124.8} = 0.055 < 1.96$$

Do not reject H_0.

Illustration.5.3.5

An educator claims that the average I.Q of American college students is at most 110 and that in a study made to test his claim, 150 American college students had an average I Q of 111,2 with a standard deviation 7.2. At 1% LOS test the claim of the educator.

Solution

$$H_0: \mu = 110$$

$$H_1: \mu > 110$$

$$|Z| = \frac{|111.2 - 110|}{\frac{7.2}{\sqrt{150}}} = 2.0412 < 2.33$$

Do not reject H_0. Claim of the educator is valid

p-value

Since the test is right tailed, $P[Z\ z] = P[Z\ 2.0412] = 0.5 - P[0 < z < 2.0412]$

$$= 0.5 - 0.4793$$

$$= 0.0207 > 0.01$$

Do not the reject null hypothesis

5.3.3 Hypothesis concerning the difference of means of two samples

To test whether the difference of means $\underline{x_1} - \underline{x_2}$ of two sample means $\underline{x_1}\ and\ \underline{x_2}$ are significant, we use the test statistics $Z = \frac{\underline{x_1} - \underline{x_2}}{SE}$ follows a standard normal distribution.

SE=$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$ where the samples n_1 and n_2 are drawn from two different populations with standard deviation σ_1 and σ_2. If the standard deviationσ_1 and σ_2 are not known,

SE=$\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$.

Illustration.5.3.6.

Ten plots of land are treated with fertiliser A and 12 with fertiliser B. The mean yield of first is 6 bushels with a stand deviation of 0.03 bushels. The yields of second plots have 5.95 bushels with a stand deviation. of 0.04 bushels. At a 1% l.o.s., is there a difference have a mean of the fertilisers?

Solution

$$n_1 = 10, \underline{x_1} = 6, s_1 = 0.03$$

$$n_2 = 12, \underline{x_2} = 5.95, s_2 = 0.04$$

$$H_0: \mu_1 = \mu_2$$

$$H_1: \mu_1 \neq \mu_2$$

S.E. of distribution of $\underline{x_1} - \underline{x_2}$ is $\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

$\alpha = 0.05, \quad d.f.\ n_1 + n_2 - 2 = 20$

$Z = \frac{\underline{x_1} - \underline{x_2}}{SE}$ where SE=$\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = SE = \sqrt{\frac{0.03^2}{10} + \frac{0.04^2}{12}} = 0.0148$

$$Z = \frac{\underline{x_1} - \underline{x_2}}{SE} = \frac{\underline{6} - 5.95}{0.0148} = \frac{0.05}{0.0148} = 3.37$$

Table value $t_{20} = 2.086$

Since $|Z| = 3.37 > 2.086$. Reject H_0

Illustration5.3.7.

The means of two random samples of size 1000 and 2000 are 67.5 and 68 inches respectively. Can the sample be regarded to have drawn from the same population of SD 9.5 inches? Test at 5%LOS

Solution

$$H_0: \mu_1 = \mu_2$$

$$H_1: \mu_1 \neq \mu_2$$

Given $n_1 = 1000,\ n_2 = 2000, \underline{x_1} = 67.5,\ \underline{x_2} = 68, \sigma = 9.5$

The test statistic is $Z = \frac{\underline{x_1} - \underline{x_2}}{SE}$

Since σ is given $SE = \sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$

$$= 9.5\sqrt{\frac{1}{1000} + \frac{1}{2000}} = 0.367$$

$$Z = \frac{\underline{x_1} - \underline{x_2}}{SE} = \frac{67.5 - 68}{0.367} = -\frac{0.5}{0.367} = -1.36$$

Since $|Z| = 1.36 < 1.96$, H_0 is accepted.

Test of significance with respect to Proportion

When a random sample of n observations are drawn from a population which has a proportion of success of p. The formula for applying the test of significance with respect to proportion is discussed hereunder.

To test whether the difference between sample proportion $\underline{p}$ and population proportion p is significant, we use the test statistic

$Z = \frac{\underline{p} - p}{SE}$ which follows a standard normal distribution. SE=$\sqrt{\frac{pq}{n}}$ where $q = 1 - p$

Illustration5.3.8.

A coin is tossed 10.000 times and head turns up 5195 times. Is the coin unbiased?

Solution

If the coin unbiased, $p = \frac{1}{2} = 0.5$

$$n = 10{,}000, \quad \underline{p} = \frac{5195}{10000} = 0.5195$$

$$H_0: \underline{p} = 0.5$$

$$H_1: \underline{p} \neq 0.5$$

$$\text{SE} = \sqrt{\frac{pq}{n}} = \sqrt{\frac{0.5 \times 0.5}{10000}} = 0.005$$

$$Z = \frac{\underline{p} - p}{SE} = \frac{0.5195 - 0.5}{0.005} = 3.9$$

Since $|Z| = 3.9 > 1.96$, H_0 is rejected. The coin is biased.

Illustration5.3.9

In a big city 325 men out of 600 men were found to be smokers. Does this information support the conclusion that the majority of men in this city are smokers? Use 5% LOS

Solution

If the coin unbiased, $p = \frac{1}{2} = 0.5$

$$n = 600, \quad \underline{p} = \frac{325}{600} = 0.5417$$

Population proportion p is taken as 1/2

$$H_0: \underline{p} = \frac{1}{2}$$

$$H_1: \underline{p} > \frac{1}{2}$$

$$\text{SE} = \sqrt{\frac{pq}{n}} = \sqrt{\frac{0.5 \times 0.5}{600}} = 0.0204$$

$$Z = \frac{\underline{p} - p}{SE} = \frac{0.5417 - 0.5}{0.0204} = 2.04$$

Since $|Z| = 2.04 > 1.645$, H_0 is rejected. The majority of men in this city are smokers.

Objective Type Questions

1. What is the first step in hypothesis testing?
2. What does the null hypothesis (H_0) represent?
3. What is the purpose of the alternative hypothesis (H_1)?
4. A two-tailed test is used when:
5. In a hypothesis test, increasing the sample size generally:
6. Define the term "population" in the context of a statistical investigation.
7. What is a sample in a statistical study?
8. What does sample size refer to in statistics?
9. State the null hypothesis for testing the mean of a population.
10. What is the significance level typically used in hypothesis testing?
11. What is the critical value for rejecting a hypothesis at the 5% level of significance using a two-tailed test?
12. When is the null hypothesis rejected at a 1% level of significance for a two-tailed test?
13. What does the test statistic represent in hypothesis testing?
14. What distribution is assumed for conducting a test of significance with respect to mean?
15. How is the Z-statistic calculated for a test of significance with respect to mean?

Answers

1. State the null and alternative hypotheses
2. The assumption that there is no effect or no difference
3. It states that there is a significant difference or effect
4. We are testing for any significant difference, regardless of direction.
5. Decreases the variability of the sample mean
6. A group of individuals under study
7. A finite subset of a population
8. The number of individuals in a sample
9. μ = a specific value
10. 0.05 or 0.01
11. ±1.96
12. If Z > ±2.58
13. A measure of how far the sample statistic is from the hypothesized population parameter
14. Normal distribution
15. $(\bar{X} - \mu) / (\sigma/\sqrt{n})$

Assignment

1. A light bulb company claims that the 100-watt light bulb it sells has an average life of 1200 hours with a standard deviation of 100 hours. For testing the claim 50 new bulbs were selected randomly and allowed to burn out. The average lifetime of these bulbs was found to be 1180 hours. Is the company's claim is true at 5% level of significance?
2. In two samples of women from Punjab and Tamilnadu, the mean height of 1000 and 2000 women are 67.6 and 68.0 inches respectively. If population standard deviation of Punjab and Tamilnadu are same and equal to 5.5 inches then, can the mean heights of Punjab and Tamilnadu women be regarded as same at 1% level of significance?
3. A machine produces a large number of items out of which 25% are found to be defective. To check this, company manager takes a random sample of 100 items and found 35 items defective. Is there an evidence of more deterioration of quality at 5% level of significance?
4. In a large population 30% of a random sample of 1200 persons had blue eyes and 20% of a random sample of 900 persons had the same blue eyes in another population. Test the proportion of blue-eyes persons is same in two populations at 5% level of significance.

5. In a sample of 100 MSc. Economics first year students of a University, it was seen that 54 students came from Science background and the rest are Large Sample Tests from other background. Can we assume that 50% of the students are from Science background in MSc. Economics first year students in the University at 1% level of significance?

Suggested reading

1. K Huffman and R Kunze, Linear Algebra, Pearson Education, 2nd Edition, 2005.
2. Thomas M. Apostol, Calculus, Wiley, 2nd Edition, 1991 ISBN 960-07-0067-2.
3. Michael Spivak. Calculus, publish or Perish, 2008, ISBN 978-0914098911.
4. Ross L. Finney, Maurice D.Weir. and Frank R. Giordano. Thomas's Calculus, Pearson l2th Edition 2009.
5. Bendat, J. S. and A. G. Piersol. Random Data: Analysis and Measurement Procedures. 4th Edition.
6. John Wiley & Sons, Inc., NY, USA, 2010 3. Montgomery, D. C. and G. C. Runger. Applied Statistics and Probability for Engineers. 5th Edition. John Wiley & Sons, Inc., NY, USA, 2011.
7. Kolman, Busby, Ross and Rehman, Discrete Mathematical Structures for Computer Science, Pearson Education,6th Edition, 2017.
8. Erwin Kreyszig, Advanced Engineering Mathematics, Wiley India, 9th Edition, 2011.

Unit 4
t test and ANOVA

Learning outcome

After completing this unit, the learner will be able to:

- Recall the purpose of a t-test and ANOVA.
- Identify when to use a t-test or ANOVA.
- State the formula for the t-statistic.
- List the steps for performing an ANOVA test.
- Recognize the assumptions of t-tests and ANOVA.

Pre-requisites

Before conducting a t-test or ANOVA, it is essential to ensure that the data meets key assumptions, including normality, independence, and homogeneity of variance. The sample data should be randomly collected and appropriately scaled (interval or ratio). For a t-test, comparisons should be between two groups, while ANOVA is used for three or more groups. Additionally, statistical software or tools should be available for calculations, and a clear hypothesis should be defined to guide the analysis.

Keywords

t test, Analysis of Variance

Discussion

The t-test and ANOVA are statistical methods used to compare group means and determine if differences are significant. A t-test is used when comparing two groups, such as testing whether the average test scores of male and female students differ significantly. On the other hand, ANOVA (Analysis of Variance) is used when comparing three or more groups, such as analyzing whether students from different schools have significantly different average scores. Both tests rely on assumptions like normality and homogeneity of variance, and they help in decision-making by determining if observed differences are due to chance or actual effects.

5.4.1. t-test

A **t-test** is a statistical test used to compare the means of two groups to determine if there is a significant difference between them. It helps assess whether the differences observed in sample data are due to actual effects or just random chance. The t-test assumes that the data is normally

distributed and that the variances of the two groups are equal (in the case of an independent t-test).

When the size of the sample is less than 30, then that sample is small sample. For small sample we use Student's t test or t test.

5.4.2.Hypothesis concerning Mean

To test whether the difference between the sample mean $\underline{x}$ and the population mean μ is significant, we use the statistic,

$$t = \frac{\underline{x}-\mu}{\frac{\sigma}{\sqrt{n}}} \text{ follows } N(0,1)$$

Where $\underline{x}$- sample mean, $\mu -$ population mean, $\sigma -$population standard deviation

n- number of observations.

If the standard deviation of the sample is given directly, then the test statistic,

$$t = \frac{\underline{x} - \mu}{\frac{s}{\sqrt{n-1}}}$$

If the calculated value of t exceeds the tabulated value of t at given level of significance, then then the null hypothesis H_0 is rejected.

If the calculated value of t is less than the tabulated value of t at given level of significance, then then the null hypothesis H_0 is accepted.

5.4.3.Student t test for difference of Mean

There are situations where the distribution of difference in means. Then we need to adopt the hypothesis tests for difference.

If the population standard deviation is not known, we use the test statistic

$$t = \frac{\underline{x_1} - \underline{x_2}}{SE}$$

Which follows t – distribution with $n_1 + n_2 - 2$ degree of freedom, where

$$\text{SE} = \sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

If σ is not known, we may assume that $\sigma = \frac{\sqrt{n_1 s_1^2 + n_2 s_2^2}}{n_1 + n_2 - 2}$

Illustration 5.3.1

A machinist is expected to make engine parts with axil diameter of 1.75 cm. A random sample of 10 parts shows a mean diameter 1.85 cm with a SD of 0.1 cm. On the basis of the sample , would you say that the work of the machinist is interior? Is the claim acceptable at 5% LOS.

Solution

Here, $n = 10, \underline{x} = 1.85, s = 0.1, \ \mu = 1.75$

$$H_0: \underline{x} = \mu$$

$$H_1: \underline{x} \neq \mu$$

Two tailed test is used.

$$t = \frac{\underline{x} - \mu}{\frac{s}{\sqrt{n-1}}}$$

$$t = \frac{1.85 - 1.75}{\frac{0.1}{\sqrt{10-1}}}$$

$$= \frac{0.1}{\frac{0.1}{3}} = 3$$

Degree of freedom = $n - 1 = 10 - 1 = 9$

From t table with degree of freedom $= 9$ $t_{0.05} = 2.26$

$$|t| = 3 > 2.26$$

Therefore, H_0 is rejected and H_1 is accepted.

Illustration 5.3.2

The mean life of a sample of 25 bulbs is found as 1550 hours and SD of 120 hrs. The company manufacturing the bulbs claims that the average life of their bulbs is 1600 hrs. Is the claim acceptable at 5% LOS.

Solution

Here, $n = 25, \underline{x} = 1550,\ s = 120,\ \mu = 1600$

$$H_0: \underline{x} = \mu$$

$$H_1: \underline{x} < \mu$$

One tailed test is used.

$$t = \frac{\underline{x} - \mu}{\frac{s}{\sqrt{n-1}}}$$

$$t = \frac{1550 - 1600}{\frac{120}{\sqrt{25-1}}}$$

$$= \frac{-50 \times \sqrt{24}}{120} = -2.04$$

Degree of freedom = $n - 1 = 25 - 1 = 24$

From t table with $\vartheta = 24$ $t_{0.05} = 1.71$

$$|t| = 2.04 > 1.71$$

Therefore H_0 is rejected and H_1 is accepted.

Illustration 5.3.3

The average number of articles produced by two machines per day are 200 and 250 with standard deviation 20 and 25 respectively on the basis of records of 25 days production. Can you regard both the machines equally effective at 1% LOS.

Solution

$$H_0: \mu_1 = \mu_2$$

$$H_1: \mu_1 \neq \mu_2$$

Here $n_1 = 25,\ n_2 = 25, \underline{x_1} = 200, \underline{x_2} = 250, s_1 = 20,\ s_2 = 25$

So that $\sigma = \sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}} = \sqrt{\frac{25 \times 20^2 + 25 \times 25^2}{25+25-2}} = \sqrt{533.85} = 23.1$

The SE$= \sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = 23.1\sqrt{\frac{1}{25} + \frac{1}{25}}=$

The test statistic $t = \frac{\underline{x_1} - \underline{x_2}}{SE}$

$= \frac{200-250}{SE}$=-7.65

From t table with $\vartheta = 25 + 25 - 2 = 48\ t_{0.01} = 2.58$

$$|t| = 7.65 > 2.58$$

Therefore H_0 is rejected and H_1 is accepted.

Illustration 5.3.4

The mean height and SD height of 8 randomly chosen soldiers are 166.9 and 8.29 cm respectively. The corresponding vales of 6 randomly chosen sailors are 170.3 and 8.5 cm respectively. Based on this data, can we conclude that soldiers are, in general, shorter than sailors?

Solution

$$H_0: \underline{x_1} = \underline{x_2}$$

$$H_1: \underline{x_1} \neq \underline{x_2}$$

Here $n_1 = 8,\ n_2 = 6, \underline{x_1} = 166, \underline{x_2} = 170.3, s_1 = 8.29,\ s_2 = 8.5$

So that $\sigma = \sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}} = \sqrt{\frac{8 \times 8.29^2 + 6 \times 8.5^2}{8+6-2}} = \sqrt{\frac{983}{12}}$

The SE= $\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} = \sqrt{\frac{983}{12}}\sqrt{\frac{1}{8}+\frac{1}{6}}$=

The test statistic $t = \frac{x_1 - x_2}{SE}$

$= \frac{200-250}{SE}$=-0.695

From t table with $\vartheta = 8 + 6 - 2 = 12$ $t_{0.01} = 1.78$

$|t| = 0.695 < 1.78$

Therefore H_0 is accepted and H_1 is rejected.

5.4.4 ANOVA (Analysis of Variance)

ANOVA is a technique that will enable us to test the significance of the difference among more than two sample means.

Analysis of variance is useful, for example, for determining

i)which of the various training methods produces the fastest learning record

ii)whether the effects of some fertilizers on the yields are significantly different

iii)whether the mean qualities of outputs of various machines differ significantly etc.

This technique finds application in nearly every type of experimental design in natural science as well as in social science.

We assumed that the two populations from which the samples were drawn had the same variance. In many situations there is a need to test the significance of differences among three or more sampling means, or equivalently to test the null hypothesis that the sample means are all equal.

The steps to perform the one way ANOVA test are given below:

- Calculate the mean for each group x.
- Calculate the average mean $\underline{X}$. This is done by adding all the means and dividing it by the total number of means.
- Find the difference between the means of the various samples and grand mean.
- Calculate the $SSC = \sum_{j=1}^{k} \quad n_j(\underline{x} - \underline{X})^2$
- Divide SSC by the degree of freedom. degree of freedom in this case is the less than the number of samples. $MSC = \frac{SSC}{degree\ of\ freedom}$
- **Find the deviations of the various items in the sample from the mean value of the sample. Do this for all samples.**
- SSE= Sum of squares of all these deviations.
- MSE= SSE/d.f. here degree of freedom is the total number of items minus the number of samples. MSE is the variance within the samples
- Compute the F ratio = $\frac{MSC}{MSE}$

- Using the f table for the specified level of significance, α, find the critical value. This is given by $F(\alpha, df_1. df_2)$.
- If F ratio > F then reject the null hypothesis.

Illustration.5.3.5

Using the following data, perform a one way analysis of variance using α = .05.

$$[\ Group1\ \ 51\ 45\ 33\ 45\ 67\ ][\ Group2\ \ 23\ 43\ 23\ 43\ 45\ ][\ Group3\ \ 56\ 76\ 74\ 87\ 56\ ]$$

Solution

Group I

value	Mean	Deviations	Sq. deviation
51	48.2	2.8	7.84
45	48.2	-3.2	10.24
33	48.2	-15.2	231.04
45	48.2	-3.2	10.24
67	48.2	18.8	353.44
			612.8

Group II

value	Mean	Deviations	Sq. deviation
23	35.4	-12.4	153.76
43	35.4	7.6	57.76
23	35.4	-12.4	153.76
43	35.4	7.6	57.76
45	35.4	9.6	92.16
			515.2

Group III

value	Mean	Deviations	Sq. deviation
56	69.8	-13.8	190.44
76	69.8	6.2	38.44
74	69.8	4.2	17.64

87	69.8	17.2	295.84
56	69.8	-13.8	190.44
			732.8

Sample means (x) for the groups:

$$48.2, \quad 35.4, \quad 69.8$$

Total Mean = 153.4

Average Mean =$\frac{153.4}{3} = 51.13$

SSB = $\sum \quad n_j(\underline{X} - \underline{x})^2 = 5 \times (48.2 - 51.13)^2 + 5 \times (35.4 - 51.13)^2 + 5 \times (69.8 - 51.13)^2 = 3022.9335$

SSE=$\sum \quad \sum \quad \left(X - \underline{X}\right)^2 = 612.8 + 515.2 + 732.8 = 1860.8$

$$MSB = \frac{SSB}{k-1} = \frac{3022.9335}{3-1} = 1511.47$$

$$MSE = \frac{SSE}{k-1} = \frac{1860.8}{15-3} = 155$$

$$F = \frac{MSB}{MSE} = \frac{1511.47}{155} = 9.75$$

Anova Table

Source of Variation	Sum of squares	Degree of freedom	Mean square	F ratio
Between samples	SSC=3022.9335	2	MSC=SSC/df=1511.47	F=$\frac{MSC}{MSE} =$ 9.75
With in samples	SSE = 1860.8	12	$MSE = \frac{SSE}{d.f} = 155$	
Total	226	19		

$$F_{critical}\,(2,12) = 3.89$$

$$F_{critical}\,(2,12) < F$$

Sample came from the same population

Illustration.5.3.6

A common test was given to a number of students taken at random from a particular class of the four departments concerned to assess the significance of possible variations in performance. Make an analysis of variance in the following

Departments			
C	M	E	I
9	12	17	13
10	13	17	12
13	11	15	12
9	14	9	18
9	5	7	15

Solution

Group I

Value	Mean	Deviations	Sq. deviation
9	10	-1	1
10	10	0	0
13	10	3	9
9	10	-1	1
9	10	-1	1
			12

Group II

Value	Mean	Deviations	Sq. deviation
12	11	1	1
13	11	2	4
11	11	0	0
14	11	3	9
5	11	-6	36
11			50

Group III

Value	Mean	Deviations	Sq. deviation
17	13	4	16

17	13	4	16
15	13	2	4
9	13	-4	16
7	13	-6	36
13			88

Group IV

Value	Mean	Deviations	Sq. deviation
13	14	-1	1
12	14	-2	4
12	14	-2	4
18	14	4	16
15	14	1	1
14			26

Sample means (x) for the groups:

$$10, \quad 11, \quad 13, \quad 14$$

Total Mean = 48

Average Mean =$\frac{48}{4} = 12$

SSC = $\sum \; n_j(\underline{X} - \underline{x})^2 = 5 \times (10-12)^2 + 5 \times (11-12)^2 + 5 \times (13-12)^2 + 5 \times (14-12)^2 = 50$

SSE=$\sum \quad \sum \quad \left(X - \underline{X}\right)^2 = 12 + 50 + 88 + 26 = 176$

$$MSC = \frac{SSC}{k-1} = \frac{50}{4-1} = \frac{50}{3} = 16.67$$

$$MSE = \frac{SSE}{n-k} = \frac{176}{20-4} = \frac{176}{16} = 11$$

$$F = \frac{MSC}{MSE} = \frac{16.67}{11} = 1.52$$

Anova Table

Source of Variation	Sum of squares	Degree of freedom	Mean square	F ratio

Between samples	SSC=50	3	MSC=SSC/df=16.67	F=$\frac{MSC}{MSE}$ = 1.52
With in samples	SSE = 176	16	$MSE = \frac{SSE}{d.f} = 11$	
Total	226	19		

$F_{critical}\,(3,16) = 3.24$, F ratio = 1.52

F ratio < $F_{critical}\,(2,12)$

Difference in sample mean is not significant. The sample could have come from the same population.

Illustration.5.3.7

Three different machines are used for a production. On the basis of the out puts, set up an ANOVA table and test whether the machines are equally effective. Given that the value of F at 5% LOS for (2,9) d.f is 4.26

Outputs		
Machine I	Machine II	Machine III
10	9	20
15	7	16
11	5	10
10	6	14

Solution

Group I

Value	Mean	Deviations	Sq. deviation
10	11.5	-1.5	2.25
15	11.5	3.5	12.25
11	11.5	-0.5	0.25
10	11.5	-1.5	2.25
10			17

Group II

Value	Mean	Deviations	Sq. deviation

9	6.75	2.25	5.0625
7	6.75	0.25	0.0625
5	6.75	-1.75	3.0625
6	6.75	-0.75	0.5625
6.75			8.75

Group III

Value	Mean	Deviations	Sq. deviation
20	15	5	25
16	15	1	1
10	15	-5	25
14	15	-1	1
15			52

$$H_0: \mu_1 = \mu_2$$

$$H_0: \mu_1 \neq \mu_2$$

Sample means (x) for the groups:

$$10, \quad 6.75, \quad 15,$$

Total Mean = 31.75

Average Mean =$\frac{31,75}{3} = 10.58$

SSC = $\sum \quad n_j(\underline{X} - \underline{x})^2 =$

$$4 \times (10 - 10.58)^2 + 4 \times (6.75 - 10.58)^2 + 4 \times (15 - 10.58)^2 = 138.1668$$

SSE=$\sum \quad \sum \quad (X - \underline{X})^2 = 17 + 8.75 + 52 = 77.75$

$$MSC = \frac{SSC}{k-1} = \frac{138.1668}{3-1} = \frac{138.1668}{2} = 69.0834$$

$$MSE = \frac{SSE}{n-k} = \frac{77.75}{12-3} = \frac{77.75}{9} = 8.639$$

$$F = \frac{MSC}{MSE} = \frac{69.0834}{8.639} = 7.99$$

Anova Table

Source of Variation	Sum of squares	Degree of	Mean square	F ratio

		freedom		
Between samples	SSC=138.1668	2	MSC=SSC/df=69.0834	F=$\frac{MSC}{MSE}$ = 7.99
With in samples	SSE = 77.75	9	$MSE = \frac{SSE}{d.f}$ = 8.639	
Total		19		

$F_{critical}\,(2,9) = 4.26$, F ratio = 7.99

F ratio > $F_{critical}\,(2,9)$

Machines are not equally effective.

Recap

- A **t-test** is a statistical test used to compare the means of two groups to determine if there is a significant difference between them.
- ANOVA is a technique that will enable us to test the significance of the difference among more than two sample means.

Objective Type Questions

1. What is the main purpose of a t-test?
2. When should an ANOVA test be used instead of a t-test?
3. When is a paired t-test used?
4. What is the assumption in the null hypothesis in a t-test
5. What is the use of one-way ANOVA test?
6. What are the key assumptions of a t-test?
7. How do you determine whether to accept or reject the null hypothesis in a t-test?
8. What is the formula for the t-test statistic when the population standard deviation is known?
9. What is the formula for the t-test statistic when the sample standard deviation is used?
10. What is the test statistic formula for comparing two sample means in a t-test?
11. What does the F-ratio represent in ANOVA?
12. How is the mean square between groups (MSC) calculated in ANOVA?
13. How is the mean square within groups (MSE) calculated in ANOVA?
14. What is the formula for computing the F-ratio in ANOVA?
15. How do you determine the critical value for the F-test in ANOVA?

Answers

1. To compare the means of two groups
2. When comparing three or more sample means
3. The same subjects are tested before and after an intervention
4. The population means are equal
5. to compare the means of more than two groups

6. Data should be normally distributed, have equal variance, and be independent.

7. Compare the calculated t-value with the table t-value at a chosen significance level.

8. The formula for a t-test is $t = \frac{x-\mu}{\sigma/\sqrt{n}}$.

9. If the sample standard deviation is given, the formula is $t = \frac{x-\mu}{s/\sqrt{n-1}}$.

10. To compare two group means, use $t = \frac{x_1-x_2}{SE}$, where SE is the standard error.

11. The F-ratio is found by dividing variance between groups by variance within groups.

12. The mean square between groups (MSC) is found by dividing sum of squares between groups by its degrees of freedom.

13. The mean square within groups (MSE) is found by dividing sum of squares within groups by its degrees of freedom.

14. The F-value is calculated as $F = \frac{MSC}{MSE}$.

15. Compare the F-value with the F-table value to check significance.

Assignment

1. A manufacturer claims that a special type of projector bulb has an average life 160 hours. To check this claim an investigator takes a sample of 20 such bulbs, puts on the test, and obtains an average life 167 hours with standard deviation 16 hours. Assuming that the life time of such bulbs follows normal distribution, does the investigator accept the manufacturer's claim at 5% level of significance?
2. The expected lifetime of electric light bulbs produced by a given process was 1500 hours. To test a new batch a sample of 10 was taken which showed a mean lifetime of 1410 hours. The standard deviation is 90 hours. Test the hypothesis that the mean lifetime of the electric light bulbs has not changed, using a level of significance of $\alpha = 0.05$.
3. The means of two random samples of sizes 10 and 8 drawn from two normal populations are 210.40 and 208.92 respectively. The sum of squares of the deviations from their means is 26.94 and 24.50 respectively. Assuming that the populations are normal with equal variances, can samples be considered to have been drawn from normal populations having equal mean.
4. Three different traffic routes are tested for mean driving time. The entries in the table are the driving times in minutes on the three different routes. perform a one way analysis of variance using $\alpha = .05$.

Route 1	Route II	Route III
30	**27**	**16**
32	**29**	**41**
27	**28**	**22**
35	**36**	**31**

Suggested reading

1. K Huffman and R Kunze, Linear Algebra, Pearson Education, 2nd Edition, 2005.
2. Thomas M. Apostol, Calculus, Wiley, 2nd Edition, 1991 ISBN 960-07-0067-2.
3. Michael Spivak. Calculus, publish or Perish, 2008, ISBN 978-0914098911.
4. Ross L. Finney, Maurice D.Weir. and Frank R. Giordano. Thomas's Calculus, Pearson l2th Edition 2009.
5. Bendat, J. S. and A. G. Piersol. Random Data: Analysis and Measurement Procedures. 4th Edition.
6. John Wiley & Sons, Inc., NY, USA, 2010 3. Montgomery, D. C. and G. C. Runger. Applied Statistics and Probability for Engineers. 5th Edition. John Wiley & Sons, Inc., NY, USA, 2011.
7. Kolman, Busby, Ross and Rehman, Discrete Mathematical Structures for Computer Science, Pearson Education,6th Edition, 2017.
8. Erwin Kreyszig, Advanced Engineering Mathematics, Wiley India, 9th Edition, 2011.

Non parametric testing

UNIT 1

Chi-Square Test of Independence Introduction

Learning outcome

After completing this unit, the learner will be able to:

- differentiate non-parametric and parametric tests
- Recall various non parametric tests
- Recall test statistics used various non parametric tests

Pre-requisites

The learner should have thorough knowledge about testing of hypothesis studied in block 4. Also have should have good idea on terms like large sample, small sample, sample size, parameter, sampling distri- butions, Z test, t test etc.

In most of tests of hypotheses in block V require various assump- tions on the distribution of the population from which Samples are drawn. For example in large sample tests like testing of mean, pro- portion etc, the Sampling distribution is assumed to be normal. In many practical situations, such assumptions may not be justified or in many cases where population may be highly skewed. In such situations, it is essential to devise various tests and methods that are independent ef population distributions and related parame- ters. These tests are usually called non parametric tests. One of the advantages of the non parametric tests, are that it can be used as shortcut methods for complicated tests and is applicable even in the case of non-numerical data

Key Concepts

Non parametric tests, U tests, H tests, Rank sum

Discussion

6.1.1 Chi-Square Test of Independence

Learning outcome: After learning this test learner is expected to define its purpose, understand when to use it, calculate its test statistic, interpret the results, interpret the results like whether there is a significant difference between observed and expected frequencies, test the independence.,

6.1.2 Contigency table: rxc table

Let r be the no. of rows and c be the number of columns. These tables ae called contingency table or simply rxc tables(read as r by c tables). Here the population is to classify each item with respect to two categories, based on qualitative attributes. For example, if a consumer testing service rates cars as excellent, superior, average or poor with respect to performance and appearance, Note that each car falls into one of the 16cells of a 4x4 table, called Contingency table. We shall test the null hypothesis

H_0: the RV's represented by the 2 classifications are independent Alternative hypothesis is

H_1: Two RV's are not independent

The test statistic is $X^2 = \Sigma r_i$

$$: \quad c \qquad j \qquad = \qquad 1(Oij - eij \qquad)^2$$

$$\frac{}{Eij}$$

Example

A company conducted a training programme for employees and collect the folowing data

Table 1: Performance in training program

	Below Average	Average	Above Average
Poor	23	60	29
Average	28	79	60
Very Good	9	49	63

Check whether there is a relationship between employees performance in the training programme and their success in the job at 0.01 significance level.

Answer

	Below Average	Average	Above Average	Total
Poor	23	60	29	112
Average	28	79	60	167
Very Good	9	49	63	121
Total	60	188	152	400

H_o : Performance in training Programme and success in job are indecent- dent.
H_1:Both are dependent

$\alpha = 0.01$

Test statistic $X^2 = 20.179$

degrees of freedom = (3 -1) (3 - 1) = 4

2 for 4 degrees of freedom is 13.277(refer table value)

0.0

1

Therefore $X^2 > X^2$
0.

Conclusion: Reject H0.

Objective type questions:

1. A researcher has obtained data from a questionnaire used for teachers and senior teachers in the form of a 3X3 contingency table for anxiety and awareness levels Which statistical test will be suitable to test whether the two variables are independent?
   Answer: Chi-square test

2. Chi-squared can be used to understand the relationship between
   a. any two variables.
   b. two categorical variables.
   c. two continuous variables.
   d. one categorical and one continuous variable.

   Answer: b

3. Chi-squared is computed by first squaring the differences between
   a. observed frequencies and expected frequencies.
   b. observed frequencies and the total sample size.
   c. observed frequencies and observed percentages.

d. expected values and observed percentages.

Answer: a

4. The chi-squared distribution often has what type of skew?
   a. Left
   b. Right
   c. It depends
   d. It is not skewed

   Answer: b

5. Which of these distributions is used for a testing hypothesis?
   a. Normal Distribution
   b. Chi-Squared Distribution
   c. Gamma Distribution
   d. Poisson Distribution

   Answer: b

6. A bag contains 80 chocolates. This bag has 4 different colors of chocolates in it. If all four colors of chocolates were equally likely to be put in the bag, what would be the expected number of chocolates of each color?
   a. 12
   b. 11
   c. 20
   d. 9

EXERCISES

1. The results of pre survey conducted 2 weeks and 4 weeks before a general election are shown in the following table.

	Two weeks before election	Four weeks before election
Part A candidate	79	91
Part B candidate	84	66
Undecided	37	43

2. A large electronics firm that hires many workers with disabilities wants to determine whether disabilities affect such workers performance. Use the level of significance 0.05 to decide on the basis of the sample data shown in the following table whether it is reasonable to maintain that the disabilities have no effect on the workers performance.

Table 2: Performance

	Above Average	Average	Below Average
Blind	21	64	17
Deaf	16	49	14
Disability	29	93	28

3. Tests of fidelity and the selectivity of 190 radio receivers produced the results shown in the following table. Use 0.01 level of significance to test whether there is a dependence between fidelity and selectivity.

Table 3: Fidelity

		Low	Average	High
Selectivity	Low	6	12	32
	Average	33	61	18
	High	13	15	0

References

1. D.C. and G.C .Runger .Applied Statistics and Probability for Engi- neers,5th Edition.John Wiley Sons ,Inc, Ny, USA 2011
2. Kolman,Busby,Ross and Rehman,Discere Mathematical structures for Computer Science ,Pearson Education,6th Edition,2017
3. Bendat,J ,John Wiley Sons ,Inc,NY,USA 2013,Montgomery.S and A.G. Piersol.Random Data: Analysis and Measurement Procedures 4th Edi- tion
4. John A Rice ,Mathematical Statistics and Data Analysis 3rd Edition, Cengage
5. Richard A .Johnson ,Miller Freund ,s Probability and Statistics for Engineers, 6th Edition PHI
6. Murray R Spiegel ,Probability and Statistics ,3rd Edition McGraw Hill Education (India) Pri

UNIT 2

Mann-Whitney or U Test

Learning outcome

After completing this unit, learners will be able to:

- differentiate non-parametric and parametric tests
- Recall various non parametric tests
- Recall test statistics used various non parametric tests
- interpret the results obtained from U test.
- identify how to calculate U statistic by ranking combined data from both groups
- interpret the result to determine if there is a statistically significant difference between the two groups based on the calculated U value and p value
- explain the key assumptions of the test

Pre-requisites

The learner should have thorough knowledge about testing of hypothesis studied in block 45. Also have should have good idea on terms like large sam- ple, small sample, sample size, parameter, sampling distributions, Z test, t test etc.

In most of tests of hypotheses in block V require various assump- tions on the distribution of the population from which Samples are drawn. For example in large sample tests like testing of mean, pro- portion etc, the Sampling distribution is assumed to be normal. In many practical situations, such assumptions may not be justified or in many cases where population may be highly skewed. In such situations, it is essential to devise various tests and methods that are independent ef population distributions and related parame- ters. These tests are usually called non parametric tests. One of the advantages of the non parametric tests, are that it can be used as shortcut methods for complicated tests and is applicable even in the case of non-numerical data.

Discussion

6.2.1 What is U-Test?

U-Test, also known as the Mann-Whitney U test, is a non-parametric test used to assess whether there is a significant difference between the distribu- tions of two independent samples. Unlike parametric tests, which assume that the data follows a normal distribution, U-Test does not make such assumptions, making it particularly useful for analyzing non-normally dis- tributed interval data. This flexibility allows researchers and data analysts to apply U-Test in a variety of fields, including psychology, medicine, and social sciences, where the data may not meet the stringent requirements of parametric tests.

6.2.2 Application of U test

U-Test is widely used in various applications of data analytics especially when researchers need to compare two independent groups. For instance, it can be employed in clinical trials to compare the effectiveness of two different treatments on patient outcomes. Additionally, U-Test is useful in market research, where analysts may want to compare customer satisfaction ratings between two different products or services. Its versatility makes it an essential tool in the arsenal of data scientists and statisticians who are tasked with drawing meaningful conclusions from empirical data.

6.2.3 Assumptions in U test

1. The two samples being compared must be independent of one another, meaning that the selection of one sample should not influence the other.

2. The data should be measured at least on an ordinal scale, allowing for meaningful ranking. Lastly, the distributions of the two groups should have a similar shape, although this assumption is less stringent than that required for parametric tests

3. Violating these assumptions can lead to misleading conclusions.

6.2.4 Advantages of Using U-Test

One of the primary advantages of U-Test is its non-parametric nature, which allows it to be applied to data that does not meet the assumptions required for parametric tests. This makes it a valuable tool for researchers dealing with real-world data that often exhibit skewness. Furthermore, U-Test is relatively simple to compute and interpret, making it accessible for practitioners who may not have training. Its ability to handle small sample sizes effectively

6.2.5 Limitations of U-Test

One notable limitation is that it only compares two groups at a time, which can be cumbersome when analyzing multiple groups. In such cases, re- searchers may need to conduct multiple U-Tests, increasing the risk of Type I errors. Additionally, while U-Test is robust to violations of normality, it may not be as powerful as parametric tests when the assumptions of those tests are met. Therefore, researchers must carefully consider the context of their data

6.2.6 Procedure

In this section we introduce two tests based on rank sums: U Test and H Test . The U test is also called Mann- Whitney Test or Wilcoxon Test. In U test , the problem is to decide whether the two populations are same or not. The test statistic is

$$Z = \frac{U_1 - \mu_{U1}}{U\,1}$$

where $U_1 = W_1 - \frac{n1(n1+1)}{n}$

OR

$$U_2 = W_2 - \frac{n_2(n_2+1)}{}$$

(Here Assume that $U_1 < U_2$; $n_1 = sizeofsample1$, $n_2 = sizeofsample2$, W_1 is the rank of the sum of sample 1 W_2 is the rank of the sum of sample 2) $\mu_{U\,1} = \frac{n_1 n_2}{}$

$$2 \qquad \frac{n_1 n_2 (n_1+n_2+1)}{}$$

$$U\,1 \qquad 12$$

H_0: Populations are identical
H_1:Populations are not identical Level of Significance :$\alpha = 0.01$
Criterion : Reject H_0 if $|Z| > 2.575$ where Z can be calculated from the 1st equation

6.2.7 Conclusion

Its non-parametric nature, ease of use, and applicability across various fields make it a staple in the toolkit of data analysts and researchers. Understand- ing the methodology, assumptions, and interpretation of U-Test results is crucial for drawing valid conclusions from empirical data.

Illustration

Check whether samples come from identical population at 0.01 level

Sample 1	Sample 2
0.63	1.13
0.17	0.54
0.35	0.96
0.49	0.26
0.18	0.39
0.43	0.88
0.12	0.92
0.20	0.53
0.47	1.01
1.36	0.48
0.51	0.89
0.45	1.07
0.84	1.11
0.32	0.58
0.40	

Answer:

STEP 1: Rank the data in increasing order of magnitude

0.12
0.18
0.20
0.26
0.32
0.35
0.39
0.40
0.43
0.45
0.47
0.48
0.51
0.53
0.54
0.58
0.63
0.84
0.88
0.89
0.92
0.96
1.01
1.07
1.11
1.13
1.36

STEP 2: Assigning ranks for each Sample

Sample 1	Sample 2
1	5
2	8
3	13
4	16
6	17
7	18
9	21
10	22
11	23
12	24
14	25
15	26
19	27
20	28
29	

STEP 3: Find Rank Sum W_1 of Sample 1 and W_2 of Sample 2

$W_1 = 162$ $W_2 = 302$

STEP 4:

$$U_1 = W_1 - \frac{n1(n1+1)}{2} = 162 - \frac{15\times16}{2} = 42$$

$$U_2 = W_2 - \frac{n2(n2+1)}{2} = 302 - \frac{14\times15}{2} = 197$$

Choose smaller Value : U_1

STEP 5:

$$\mu_{U1} = \frac{n_1 n_2}{} = \frac{15\times14}{} = 105$$

$$\sigma^2_{U1} = \frac{n_1 n_2 (n_1+n_2+1)}{12} = \frac{14\times15(15+14+1)}{12} = 525$$

STEP 6:

$$Z = \frac{U_1 - \mu_{U1}}{U1} = \frac{42 - 105}{\sqrt{5}} = -2.75 \quad |Z| = 2.75 > 2.5$$

H_0 is rejected

Exercises

1. The Following are the numbers of minutes it took a sample of 15 men and 12 women to complete the application form for a position. Use the U test at the 0.05 level of significance to test the null hypothesis that the two samples come from identical populations against the alternative that the 2 populations are no identical

en	Women
16.5	18.6
20.0	17.8
17.0	18.3
19.8	16.6
18.5	20.5
19.2	16.3
19.0	19.3
18.2	18.4
20.8	19.7
18.7	18.8
16.7	19.9
18.1	17.6
17.9	
16.4	
18.9	

2. Comparing 2 kinds of emergency flares, a consumer testing service obtained the following burning times (rounded to the nearest tenth of a minute):

Brand C	Brand D
19.4	16.5
21.5	15.8
15.3	24.7
17.4	10.2
16.8	13.5
16.6	15.9
20.3	15.7
22.5	14.0
21.3	12.1
23.4	17.4
19.7	15.6
21.0	15.8

Use the test at the 0.01 level of significance to check whether it is said that the populations of burning times of the two kinds of flares are identical

3. The following are the scores which random samples of students from 2 minority groups obtained on a current events test. Use the U test at the 0.05 level of significance to test whether or not students from the 2 minority groups can be expected to score equally well on the test.

Minority Group 1	Minority Group 2
73	51
82	42
39	36
68	53
91	88
75	59
89	49
67	66
50	25
86	64
57	18
65	76
70	74

4. The following are data on the breaking strength (in pounds) of 2 kinds of material Use the U test at the 0.05 level of significance to test the claim that the strength Material 1 is stochastically larger than the strength of Material 2.

Material 1	Material 2
144	175
181	164
200	172
187	194
169	176
171	198
186	154
194	134
176	169
182	164
133	185
183	159
197	161
165	189
180	198
198	164

References

1. D.C. and G.C .Runger .Applied Statistics and Probability for Engi- neers,5th Edition.John Wiley Sons, Inc, Ny, USA 2011.
2. Kolman,Busby,Ross and Rehman,Discere Mathematical structures for Computer Science, Pearson Education, 6th Edition,2017.
3. Bendat,J, John Wiley Sons, Inc,NY,USA 2013, Montgomery.S and A.G. Piersol.Random Data: Analysis and Measurement Procedures 4th Edi- tion.
4. John A Rice, Mathematical Statistics and Data Analysis 3rd Edition, Cengage.
5. Richard A .Johnson, Miller Freund,s Probability and Statistics for Engineers, 6th Edition PHI.
6. Murray R Spiegel, Probability and Statistics, 3rd Edition McGraw Hill Education (india) Private Limited.

Unit 3

Wilcoxon Signed -Rank test

Learning outcome

After completing this unit, learners will be able to:

- differentiate non-parametric and parametric tests
- Recall various non parametric tests
- Recall test statistics used various non parametric tests
- determine whether there is a statistically significant difference between two related groups
- analyze magnitude and direction of the differences between paired ob- servations

Pre-requisites

Wilcoxon Signed rank test is a non-parametric rank test to compare two paired samples, whether values in one are bigger than in the other. Wilcoxon signed-rank test, also known as Wilcoxon matched pair test is that compares the median of two paired groups and tells if they are identically distributed or not In this test differences between the two variables are ranked. The null hypothesis is that the mean of these ranks is zero, which implies that the distribution of differences in the population is symmetric about zero. The alternative hypothesis is that it is not so, and thus values of one of the variables tend to be higher.

The learner should have thorough knowledge about testing of hypothesis studied in block 4. Also have should have good idea on terms like large sam- ple, small sample, sample size, parameter, sampling distributions, Z test, t test etc.

In most of tests of hypotheses in block V require various assump- tions on the distribution of the population from which Samples are drawn. For example in large sample tests like testing of mean, pro- portion etc, the Sampling distribution is assumed to be normal. In many practical situations, such assumptions may not be justified or in many cases where population may be highly skewed. In such situations, it is essential to devise various tests and methods that are independent ef population distributions and related parame- ters. These tests are usually called non parametric tests. One of the advantages of the non parametric tests, are that it can be used as shortcut methods for complicated tests and is applicable even in the case of non-numerical data.

Discussion

6.3.1 Things to Remember

Wilcoxon Signed Rank test is used

- with matched/paired data.
- Null Hypothesis: The median difference between the pairs is zero.
- W: Test Statistic Non-zero median differences will cause this value to be very large (negative or positive Small values when median difference is zero/close to zero.
- Null Hypothesis: The median difference between the pairs is zero.
- Non-zero median differences will cause this value to be very large (neg- ative or positive)
- Small values when median difference is zero/close to zero.
- p-value: significance value used to determine if null hypothesis should be rejected (i.e. p <0.05 -> reject)
- significance value used to determine if null hypothesis should be rejected
- The Wilcoxon Signed Rank Test is the non-parametric version of the paired t-test. It is used to test whether or not there is a significant difference between two population means

- Use the Wilcoxon Signed Rank test when we would like to use the paired t-test but the distribution of the differences between the pairs is severely not normally distributed.

Exercises

A basketball coach wants to know if a certain training program increases the number of free throws made by his players. To test this, he has 15 players shoot 20 free throws each before and after the training program.

player	before	after
1	14	15
2	17	17

3	12	15
4	15	15
5	15	17
6	9	14
7	12	9
8	13	14
9	13	11
10	15	16
11	19	18
12	17	20
13	14	20
14	14	10
15	16	17

Since each player can be "paired" with themselves, he had planned on us- ing a paired t-test to determine if there was a significant difference between the mean number of free throws made before and after the training program. The coach uses Wilcoxon Signed Rank Test as shown below.

The following table gives the number of free throws made (out of 20 at- tempts) by each of the 15 players, both before and after the training program:

Null hypothesis H0: The median difference between the two groups is zero. Alternate hypothesis H1: The median difference is negative. (e.g. the players make less free throws before participating in the training program).

player	Before	After	Rank difference	Absolute difference	Rank
2	17	17	0	0	-
4	15	15	0	0	-
1	14	15	-1	1	3
8	13	14	-1	1	3
10	15	16	-1	1	3
11	19	18	1	1	3
15	16	17	-1	1	3
5	15	17	-2	2	6.5
9	13	11	2	2	6.5
3	12	15	-3	3	9
7	12	9	3	3	9
12	17	20	-3	3	9
14	14	10	4	4	11
6	9	14	-5	5	12
13	14	20	-6	6	13

- Order the pairs by the absolute differences and assign a rank from the smallest to largest absolute differences
- While ordering Ignore pairs that have an absolute difference of "0" and assign mean ranks when there are ties.
- Find the sum of the positive ranks= 29.5 and the negative ranks = -61.5.
- The test statistic, W, is the smaller of the absolute values of the pos- itive ranks and negative ranks. Here the smaller value is 29.5. Thus, the test statistic is W = 29.5.
- Choose alpha 0.05. From table critical value =17 for n = 13 (two pairs we discard as they have zero difference)
- Conclusion: Since 29.5 > 17, accept Ho. That is we do not have sufficient evidence to say that the training program leads to a significant increase in the number of free throws made by the players.

6.3.2 Multiple Choice Questions

1. Select all of the following statements which you believe to be True about the Wilcoxon signed-ranks test

   a. It assumes that the pairs of values do not follow a Normal distri- bution
   b. It makes no assumptions about the distribution of the data.
   c. It is equivalent to the Mann-Whitney U test when the number of observations is equal.
   d. Is also known as the Wilcoxon Rank-Sum Test

   Answer b

2. The best use of Wilcoxon signed-ranks test will be for comparison of which of the following types of data.

   a. Continuous, parametric unpaired data
   b. Continuous, non-parametric paired data
   c. Continuous, non-parametric, unpaired data
   d. Categorical unpaired data
   e. Continuous, parametric paired data

   Answer: b

3. A researcher measured the same group of people's physiological reac- tions while watching horror films and compared them to when watching erotic films. The resulting

data were skewed. What test should be used to analyse the data?

a. Independent t-test
b. Wilcoxon signed-rank test
c. Dependent (related) t-test
d. Mann–Whitney test

Answer: b

4. What is the value of the test statistic W for the above data ? There are 11 data for version one and two. The data is given above.

version	1	2	3	4	5	6	7	8	9	10	11
1	2	7	7	10	19	20	21	13	12	20	25
2	3	7	9	14	14	13	15	21	21	23	25
diff	1	0	2	4	-5	-7	-6	8	9	3	0

Answer: 18

References

6. D.C. and G.C .Runger .Applied Statistics and Probability for Engi- neers,5th Edition.John Wiley Sons ,Inc, Ny, USA 2011
7. Kolman,Busby,Ross and Rehman,Discere Mathematical structures for Computer Science ,Pearson Education,6th Edition,2017
8. Bendat,J ,John Wiley Sons ,Inc,NY,USA 2013,Montgomery.S and A.G. Piersol.Random Data: Analysis and Measurement Procedures 4th Edi- tion
9. John A Rice ,Mathematical Statistics and Data Analysis 3rd Edition, Cengage
10. Richard A .Johnson ,Miller Freund ,s Probability and Statistics for Engineers, 6th Edition PHI
11. Murray R Spiegel ,Probability and Statistics ,3rd Edition McGraw Hill Education (india) Private Limited .

Unit 4

Kruskal - Wall's Test or H Test

Learning outcome

After completing this unit, learners will be able to:

- differentiate non-parametric and parametric tests
- Recall various non parametric tests
- Recall test statistics used various non parametric tests
- determine statistically significant difference between two or more groups of an independent variable on a continuous or ordinal dependent vari- able

Pre-requisites

The learner should have thorough knowledge about testing of hypothesis studied in block 4. Also have should have good idea on terms like large sam- ple, small sample, sample size, parameter, sampling distributions, Z test, t test etc.

In most of tests of hypotheses in block V require various assump- tions on the distribution of the population from which Samples are drawn. For example in large sample tests like testing of mean, pro- portion etc, the Sampling distribution is assumed to be normal. In many practical situations, such assumptions may not be justified or in many cases where population may be highly skewed. In such situations, it is essential to devise various tests and methods that are independent of population distributions and related parame- ters. These tests are usually called non parametric tests. One of the advantages of the non parametric tests, are that it can be used as shortcut methods for complicated tests and is applicable even in the case of non-numerical data.

Keywords

Discussion

The test is named after the scientists who discovered it, William Kruskal and W. Allen Wallis. The major purpose of the test is to check if the sample is tested if the sample is taken from the same population or not.Sometimes referred to as a one way ANALYSIS OF VARIANCE(ANOVA) on ranks, Kruskal Wall or H test is a non-parametric test that is used

to determine the statistical differences between the two or more groups of an independent variable. Kruskal Wall is the generalisation of the U test. The test is used to test the null hypothesis that k independent random samples come from identical populations

i. Key points in H test

- Non parametric
- No assumption of normality or equal variance
- It compares ranks
- More robust to outliers and non-normal data

ii. Procedure

As in the U test all the observations are ranked jointly. If R_i is the sum of the ranks occupied by the n_i observations of the i^{th} sample and $n_1 + n_2 + n_3 + ... + n_k = n$, the test is based on the statistic

$$H = \frac{12}{} \sum$$

$$R^2$$

$$i \quad - 3(n + 1) \qquad —$$

$$n(n+1) \; _{i=1} \; n_i$$

when $n_i > 5$ for all i and the null hypothesis is true, the sampling distribution of the H statistic is well approximated by the Chi - Square distribution with K-1 degree of freedom

iii. Illustration

An Experiment designed to compare three preventive method against corrosion yielded the following maximum depths of the pits (in thousandths of an inch) in pieces of wire subjected to the respective treatments.

Method A	77	54	67	74	71	66	
Method B	60	41	59	65	62	64	52
Method C	49	52	69	47	56		

Use the 0.05 level of significance to test the null hypothesis that the three samples come from identical populations

Solution:

H_0: Populations are identical

H_1:Populations are not equal Level of Significance :$\alpha = 0.05$

Criterion : Reject H_0 if $H > 5.991$ (from X^2*table* for $\alpha = 2$ degrees freedom) Calculation of

Method A	6	13	14	16	17	18	
Method B	1	4.5	8	9	10	11	12
Method C	2	3	4.5	5	7	15	

$$R_1 = 84$$
$$R_2 = 55.5$$
$$R_3 = 31.5$$
$$H = \frac{12}{18\ 19\ 6\ 7\ 5}\left(\frac{84^2}{} + \frac{55.5^2}{} + \frac{31.5^2}{}\right) - 3.19$$
$$H = 6.7$$

Therefore H_0 is rejected
Three preventative methods against corrosion are not equally effective

Example 1: Comparing Study Techinques You randomly split up a class of 90 students into three groups of 30. Each group uses a different studying technique for one month to prepare for an exam. At the end of the month, all of the students take the same exam. You want to know whether or not the studying technique has an impact on exam scores. From previous studies you know that the distributions of exam scores for these three studying tech- niques are not normally distributed so you conduct a Kruskal-Wallis test to determine if there is a statistically significant difference between the median scores of the three groups.

Example 2: Comparing Sunlight Exposure You want to know whether or not sunlight impacts the growth of a certain plant, so you plant groups of seeds in four different locations that experience either high sunlight, medium sunlight, low sunlight or no sunlight. After one month you measure the height of each group of plants. It is known that the distribution of heights for this certain plant is not normally distributed and is prone to outliers. To determine if sunlight impacts growth, you conduct a Kruskal-Wallis test to determine if there is a statistically significant difference between the median height of the four groups.

Exercise of Kruskal-Wallis Test

A researcher wants to know whether or not three drugs have different ef- fects on knee pain, so he recruits 30 individuals who all experience similar knee pain and randomly splits them up into three groups to receive either Drug 1, Drug 2, or Drug 3. After one month of taking the drug, the re- searcher asks each individual to rate their knee pain on a scale of 1 to 100, with 100 indicating the most severe pain. The ratings for all 30 individuals are shown below:

Drug1	Drug2	Drug3
78	71	57
65	66	88
63	56	58
44	40	78
50	55	65
78	31	61
70	45	62
61	66	44
50	47	48
44	42	77

The researcher wants to know whether or not the three drugs have dif- ferent effects on knee pain, so he conducts a Kruskal-Wallis Test using a .05 significance level to determine if there is a statistically significant difference between the median knee pain ratings across these three groups.
The null hypothesis (H0): The median knee-pain ratings across the three groups are equal Since the p-value of the test (0.21342) is not less than 0.05, we accept the null hypothesis.

Conclusion

Parametric tests Vs Non parametric test

In block 5 we have studied large sample tests like testing the mean/means, z tests etc and small sample tests like t -test. These tests are usually called parametric tests. In block 6 we mainly concentrates on non parametric tests. Here we shall compare major differences in parametric tests and non para- metric tests.

After an understanding of the two broad categorization of the statistical tests it is easier to differentiate between them. This differentiation is important in the light that their justified application to the data brings increased validity to the findings.

Although parametric tests has greater popularity in the field of statistical applications, the non-parametric tests have their own utility and advantages. Let us compare the advantages and disadvantages of both in terms of appli- cations, assumptions and statistical power.

Ease of application and simplicity As the non-parametric tests are con- cerned with data in basic

scales it is easier and simpler to apply them. .The parametric tests on the other hand need mathematical knowledge and are difficult to comprehend by the researchers with no mathematical background.

Simpler or no assumptions compliance The non-parametric tests in comparison to parametric ones are based on fewer and flexible assumptions. As the compliance to a number of assumptions about the population parame- ters is not needed the non-parametric tests enjoy wider scope of applications. **Violation in assumptions** The parametric tests have rigid assumptions to adhere to and thus, are difficult to comply. These violated assumptions effect the validity of the results when the tests are used despite of the assumptions. The non parametric tests are less susceptible to violations, which can be easily detected.

Disadvantages of Non parametric tests

- Less powerful than parametric tests if assumptions haven't been vio- lated.
- More labor-intensive to calculate by hand
- Critical value tables for many tests aren't included in many computer software packages.
- Less efficient as compared to parametric test.
- The results may or may not provide an accurate answer because they are distribution free.
- Non parametric analyses might not provide accurate results when vari- ability differs between groups.

Recap

- Mann Whitney U test is used for testing the difference between two independent groups with dependent variable.
- Wilcoxon signed rank test is used for testing the difference between two related variables which takes into account the magnitude and sign of difference.
- Kruskal-Wallis test compares the outcome among more than two inde- pendent groups by making use of the medians.
- The Chi-Square Test of Independence is used to determine whether or not there is a significant association between two categorical variables

Objective-type questions

1. What are Non-Parametric Tests?
2. How are Non-Parametric tests different from Parametric tests?
3. When can I apply non-parametric tests?
4. What are the Pros and Cons of using non-parametric test?
5. What are non-parametric tests? Give examples for it.
6. Is Chi Square non-parametric test?

Suggested reading

1. Erwin Kreyszig,Advanced Engineering Mathematics,Wiley India ,10thEdi- tion
2. Jay L.Devore,Probability and statistics for Engineering and the sci- ences,8th edition, Cengage,2012.
3. B.S.Grewal,Higher Engineering Mathematics,Khanna Publishers, 36 th Edition,2010

References

1. D.C. and G.C .Runger .Applied Statistics and Probability for Engi- neers,5th Edition.John Wiley Sons ,Inc, Ny, USA 2011
2. Kolman,Busby,Ross and Rehman,Discere Mathematical structures for Computer Science ,Pearson Education,6th Edition,2017
3. Bendat,J ,John Wiley Sons ,Inc,NY,USA 2013,Montgomery.S and A.G. Piersol.Random Data: Analysis and Measurement Procedures 4th Edi- tion
4. John A Rice ,Mathematical Statistics and Data Analysis 3rd Edition, Cengage
5. Richard A .Johnson ,Miller Freund ,s Probability and Statistics for Engineers, 6th Edition PHI
6. Murray R Spiegel ,Probability and Statistics ,3rd Edition McGraw Hill Education (india) Private Limited.

സർവ്വകലാശാലാഗീതം

വിദ്യയാൽ സ്വതന്ത്രരാകണം
വിശ്വപൗരരായി മാറണം
ഗ്രഹപ്രസാദമായ് വിളങ്ങണം
ഗുരുപ്രകാശമേ നയിക്കണേ

കൂരിരുട്ടിൽ നിന്നു ഞങ്ങളെ
സൂര്യവീഥിയിൽ തെളിക്കണം
സ്നേഹദീപ്തിയായ് വിളങ്ങണം
നീതിവൈജയന്തി പാറണം

ശാസ്ത്രവ്യാപ്തിയെന്നുമേകണം
ജാതിഭേദമാകെ മാറണം
ബോധരശ്മിയിൽ തിളങ്ങുവാൻ
ജ്ഞാനകേന്ദ്രമേ ജ്വലിക്കണേ

കുരീപ്പുഴ ശ്രീകുമാർ

SREENARAYANAGURU OPEN UNIVERSITY

Regional Centres

Kozhikode
Govt. Arts and Science College
Meenchantha, Kozhikode,
Kerala, Pin: 673002
Ph: 04952920228
email: rckdirector@sgou.ac.in

Thalassery
Govt. Brennen College
Dharmadam, Thalassery,
Kannur, Pin: 670106
Ph: 04902990494
email: rctdirector@sgou.ac.in

Tripunithura
Govt. College
Tripunithura, Ernakulam,
Kerala, Pin: 682301
Ph: 04842927436
email: rcedirector@sgou.ac.in

Pattambi
Sree Neelakanta Govt. Sanskrit College
Pattambi, Palakkad,
Kerala, Pin: 679303
Ph: 04662912009
email: rcpdirector@sgou.ac.in

COMPUTATIONAL FOUNDATIONS FOR DATA SCIENCE

COURSE CODE: B24DS02DC

ISBN 978-81-984025-1-6

Sreenarayanaguru Open University
Kollam, Kerala Pin- 691601, email: info@sgou.ac.in, www.sgou.ac.in Ph: +91 474 2966841