MERITSHOT

1. Introduction to Statistics and its importance

Statistics

MERITSHOT

- Bunch of numbers, formulas, and charts.
- Academia, business, healthcare, etc.
- Statistics is the universal language,
    - helps us make informed decisions.

Statistics gives Tools

MERITSHOT

Describe and Summarize Data:

- Helps you make sense of the numbers.
- By providing
  - Data's central tendencies,
  - Variabilities, and
  - Distributions.
- Identifying patterns and trends.

Statistics gives Tools

MERITSHOT

Making Predictions:

- Statistics enables you to make educated guesses about future outcomes.
- Invaluable in fields,
    - Finance and economics.

Statistics gives Tools

MERITSHOT

Testing Hypotheses:

- If you have a hunch or a theory,
  - Statistics
    - Test it.
- Scientific research and decision-making.

Statistics gives Tools

MERITSHOT

Solve Real-World Problems:

- Statistics is not just for statisticians.
- It's a problem-solving tool,
    - applied to almost any domain.
- Helps you make sense of complex situations and guide your actions.

Why Python?

MERITSHOT

- Python, known for its
    - Simplicity and versatility.
- Especially in the field of data science.
- Ecosystem for data analysis.

Why Python is Go-To Tool for Statistics

MERITSHOT

Libraries:

- Python boasts an array of libraries,
  - NumPy, pandas, Matplotlib, and Seaborn.
- Designed for,
  - data manipulation, analysis, and visualization.
- Libraries simplify complex statistical operations.

Why Python is Go-To Tool for Statistics

MERITSHOT

Abundant Resources:

- Python has a massive online community.
- Countless resources,
    - easy to find solutions and learn from experts.

Why Python is Go-To Tool for Statistics

MERITSHOT

Integration with Data Sources:

- Integrates with various data sources,
  - Databases,
  - Spreadsheets, and
  - Web APIs,
- Ideal for gathering and cleaning data.

Why Python is Go-To Tool for Statistics

MERITSHOT

Versatility for Research:

- Allows for seamless integration with other tools and technologies.
- Creating reproducible research workflows.

2. Explain role of statistics in data analysis

Power of Data

MERITSHOT

- Digital age.
- Everything from smartphones and social media,
  - Scientific experiments and business transactions.
- Holds the potential to reveal valuable insights.
- Statistics is the art and science of
  - Collecting,
  - Organizing,
  - Analyzing,

Power of Data

MERITSHOT

    - Interpreting, and
    - Presenting data.
- Within the numbers,
- Allowing us to make informed decisions,
- Identify trends, and
- Draw meaningful conclusions.

Statistics: Data Detective

MERITSHOT

- Imagine you're a detective investigating a complex case.
- Evidence is scattered.
- Helping you gather and interpret evidence to solve the mysteries that data presents.

How Statistics Serves as Data Detective

MERITSHOT

Data Collection:

- Design surveys, experiments, and data collection methods to gather relevant information.
- Ensures that your data is,
  - Accurate,
  - Representative, and
  - Unbiased.

How Statistics Serves as Data Detective

Data Cleaning:

- With missing values, outliers, and inconsistencies.
- Techniques to clean and preprocess the data,
    - making it ready for analysis.

How Statistics Serves as Data Detective

MERITSHOT

Data Summarization:

- Statistics provides tools for summarizing data.
- Measures like mean, median, and standard deviation,
    - Help you understand the data's central tendencies and variability.

How Statistics Serves as Data Detective

MERITSHOT

Data Visualization:

- Statistics introduces the art of data visualization.
- With charts, graphs, and plots,
  - Transform numbers into visual stories.
- Patterns and trends that may remain hidden in raw data.

How Statistics Serves as Data Detective

Inference:

- Make inferences about populations based on samples.
- Draw conclusions about a larger group without surveying every member.

How Statistics Serves as Data Detective

MERITSHOT

Hypothesis Testing:

- Testing your ideas and determining,
    - Evidence supports or contradicts them.

Prediction:

- Build predictive models that forecast future trends or outcomes.
- Invaluable in fields like
    - Finance, marketing, and healthcare.

How Statistics Serves as Data Detective

MERITSHOT

Decision-Making:

- By providing insights into
  - Customer behavior,
  - Market trends, and
  - Product performance.
- Allocate resources and minimize risks.

How Statistics Serves as Data Detective

MERITSHOT

Quality Control:

- In manufacturing and industry,
  - Statistics plays a critical role in quality control.
- Ensures products meet standards.
- Identify defects early in the production process.

MERITSHOT

3. Introduction to Python for Statistical Analysis

Why Python?

MERITSHOT

- Python is a versatile and powerful programming language.
- Data analysis.
- Machine learning, and
- Scientific computing for several reasons.

Why Python?

MERITSHOT

Beginner-friendly:

- Clear and readable syntax makes it easy for anyone.
- Regardless of their programming background.
- Don't need to be a coding expert to use Python effectively for statistical analysis.

Why Python?

MERITSHOT

Data Analysis:

- Rich ecosystem of libraries tailored for scientific computing and data analysis.
- Libraries like NumPy, pandas, matplotlib, and Seaborn.
- Tools and functionality needed to manipulate, visualize, and analyze data.

Why Python?

MERITSHOT

- Python is versatile enough to be used across various fields,
  - Finance to healthcare,
  - And from social sciences to engineering.

Python Basics

MERITSHOT

1. Variables:

- Assign values to variables,
  - which are like containers for data.
- ***Example:*** Create a variable named `age` and set it to 30.

2. Data Types:

- Python supports various data types,
  - Integers, floating-point numbers, strings, etc.
  - Handling data in statistics.

Python Basics

MERITSHOT

3. Lists:

- Lists are collections of items that can be of any data type.
- Allow to store and manipulate data.

4. Conditional Statements:

- Use conditional statements like
  - `if`, `else`, and `elif`.
  - Make decisions in your code based on certain conditions.

Python Basics

MERITSHOT

5. Loops:

- Loops, such as `for` and `while`,
  - repeat a set of instructions multiple times.
- Iterating through datasets.

6. Functions:

- Encapsulate a block of code and reuse it.
- Helpful when performing repetitive tasks.

Python Basics

MERITSHOT

Financial Analysis:

- Portfolio management,
- Risk assessment, and
- Predicting stock prices.

Healthcare:

- Helps analyze medical data,
- Improve patient care,
- Drug discovery, and
- Epidemiological studies

Python Basics

MERITSHOT

Marketing:

- Python assists in
  - Customer segmentation,
  - Market basket analysis, and
  - Sentiment analysis for better marketing strategies.

Social Sciences:

- Analyze survey data,
- Studying social trends, and
- Understanding human behavior.

Python Basics

MERITSHOT

Environmental Science:

- Modeling climate change,
- Monitoring ecosystems, and
- Analyzing environmental data.

Business Intelligence:

- Helps organizations make data-driven decisions,
  - Analyzing customer behavior,
  - Sales trends, and
  - Operational efficiency.

MERITSHOT

1. Types of Data

What is Data?

MERITSHOT

- A collection of facts, observations, measurements, or information,
  - Can be processed or analyzed.
- Numbers, text, images, audio, or video.
- Data is fundamental as it serves as the basis for statistical analysis and inference.

Data is Relevant in Statistics for Reasons

MERITSHOT

1. Descriptive Statistics:

- Describe and summarize characteristics of interest in a population or sample.
- Provide measures,
  - Mean, median, mode, standard deviation, and percentiles,
- Understand the central tendency, variability, and distribution of the data.

Data is Relevant in Statistics for Reasons

MERITSHOT

2. Inferential Statistics:

- Making inferences or drawing conclusions about a population based on a sample.
- Inferential statistics use probability theory and sampling techniques,
    - generalize findings from a subset of data (sample) to a larger population.

Data is Relevant in Statistics for Reasons

3. Hypothesis Testing:

- Test hypotheses and determine the significance of observed differences or relationships.
- Formulating a null hypothesis, and
- An alternative hypothesis, collecting data, and assessing the evidence against the null hypothesis.

Data is Relevant in Statistics for Reasons

4. Statistical Modeling:

- Data is employed to build statistical models that represent relationships, patterns, or predictions in data.
- Models can be used to analyze
  - complex data,
  - identify trends,
  - make predictions, or
  - simulate scenarios.

Data is Relevant in Statistics for Reasons

MERITSHOT

5. Decision Making:

- Data plays a vital role in decision making.
- Insights and evidence to support or guide decisions in various fields, including
    - Business,
    - Economics,
    - Medicine,
    - Social sciences, etc.

Data is Relevant in Statistics for Reasons

MERITSHOT

- Data quality, including factors such as
  - Completeness,
  - Accuracy,
  - Consistency, and relevance,
- Influences the validity and reliability of statistical conclusions.

Different Types of Data

MERITSHOT

Numerical data:

- Continuous or discrete numerical values.

Two subtypes:

- ***Continuous Numerical Data:***
- Can take any value within a specific range.
- ***Examples:*** Age, height, weight, temperature, etc.

Different Types of Data

MERITSHOT

Discrete Numerical Data:

- Consists of whole numbers or integers.
- ***Examples:*** Number of siblings, number of pets, number of cars, etc.
- Use statistical measures such as mean, median, mode, standard deviation, variance, percentiles, etc.
- NumPy and Pandas libraries provide functions and methods to perform the calculations.

Different Types of Data

Categorical Data:

- Non-numerical or discrete values that belong to specific categories.

Two subtypes:

- ***Nominal Data:*** Represents categories that have no specific order or ranking.
- ***Examples:*** Gender, color, occupation, marital status, etc.

Different Types of Data

MERITSHOT

- ***Ordinal Data:***

  Represents categories that have a specific order or ranking.
- ***Examples:*** Education level, Customer satisfaction rating, etc.
- Calculate frequencies, proportions, and create data visualizations,
  - Bar plots,
  - Pie charts, and histograms
- Using libraries like pandas, matplotlib, and seaborn.

MERITSHOT

2. Measures of Central Tendency

Measures of Central Tendency

MERITSHOT

- ***Mean:*** Also known as the average, is calculated by summing all the values.
- Dividing by the total number of values.
- Sensitive to extreme values and provides a measure of the overall "typical" value.

Mean = (Sum of all values) / (Number of values)

Measures of Central Tendency

- ***Median:*** Median is the middle value in a dataset,
  - Sorted in ascending or descending order.
- If the dataset has an odd number of values,
  - Median is the middle value.
- If the dataset has an even number of values,
  - Median is the average of the two middle values.
- Provides a measure of "middle" value.
- To calculate the median, data must be ordered first.

Measures of Central Tendency

- ***Mode:*** Occurs most frequently in a dataset.
- Represents the value(s) with highest frequency or occurrence.
- A dataset may have one mode, two modes, or more than two modes.
- No mode if no value is repeated.
- Useful for identifying the most common value(s) in a dataset.

Measures of Central Tendency

MERITSHOT

- Describe the distribution and summarize the data.
- Insights into the typical value and help in understanding the central characteristics of a dataset.
- Based on the data type, distribution, and the purpose of analysis.

MERITSHOT

4. Measures of Dependence

Measures of Dependence

MERITSHOT

1. Data Analysis and Research:

- Explore and uncover relationships between variables.
- By quantifying the strength and direction of the relationship,
    - Basis for further analysis and investigation.
- ***For Example:***
    - Study the relationship between income and education level or the association between advertising expenditure and sales.

Measures of Dependence

2. Prediction and Forecasting:

- Assist in predictive modeling and forecasting.
- If two variables show a strong dependence, knowledge of one variable can help predict the other.
- ***For Example:*** In finance, understanding the dependence between stock prices and market indices can aid in forecasting market movements.

Measures of Dependence

- ***3. Decision-Making and Policy Formulation:***
- In decision-making processes and policy formulation.
- Understanding the relationship between different factors,
    - Assess the impact of changes in one variable.
- Informed decisions and formulating effective policies.
- ***For Example:*** Analyzing the dependence between unemployment rates and economic growth can guide policymakers in designing strategies to stimulate employment.

Measures of Dependence

4. Risk Assessment and Portfolio Management:

- Assessing and managing risks.
- By examining the dependence between different assets or variables,
    - Stocks or financial instruments,
    - Investors can construct diversified portfolios that mitigate risks.
- Helps in managing and optimizing investment strategies.

Measures of Dependence

MERITSHOT

5. Quality Control and Process Improvement:

- To identify relationships between input variables and output quality.
- Helps in process improvement, identifying critical factors, and optimizing production processes.

MERITSHOT

5. Measures of Shape and Position

Measures of Shape

1. Skewness:

- Measures the asymmetry of the data distribution.
- Identify whether the data is skewed,
  - Left (negative skewness) or
  - Right (positive skewness) or
  - Symmetric (zero skewness).
- Skewness is valuable in fields such as finance,
  - indicate the presence of market trends or imbalances.

Measures of Shape

2. Kurtosis:

- Kurtosis measures the shape of the data distribution's tails.
- Quantifies the degree of peakedness or flatness of the distribution compared to a normal distribution.
- High kurtosis indicates heavy tails,
    - useful in financial risk analysis or detecting outliers.

Measures of Position

MERITSHOT

Percentiles:

- Percentiles divide a dataset into 100 equal parts.
- Indicate the value below which a certain percentage of the data falls.
- Percentiles are helpful in analyzing income,
    - distributions, exam scores, etc.
    - where relative rankings are significant.

Measures of Position

MERITSHOT

Quartiles:

- Quartiles divide a dataset into four equal parts.
- Identify the spread and central tendency of the data distribution.
- First quartile (Q1) represents the 25th percentile,
- Second quartile (Q2) is the median, and
- Third quartile (Q3) is the 75th percentile.
- Analyzing data in areas,
    - economics, healthcare, and demographics.

Real-World Applications: Shape & Position

MERITSHOT

1. Data Analysis:

- Insights into the characteristics of data.
- Enabling analysts to understand and describe the distribution.
- Aid in identifying patterns, trends, and potential outliers.

Real-World Applications: Shape & Position

MERITSHOT

2. Risk Assessment:

- In finance and insurance,
  - Help assess risk by identifying skewed or heavy-tailed distributions.
- Extreme events or market volatility.

Real-World Applications: Shape & Position

MERITSHOT

3. Decision-Making:

- Important information for decision-making.
- Percentiles and quartiles help establish benchmarks.
- Assess the performance of individuals or entities relative to the population.

Real-World Applications: Shape & Position

4. Comparisons:

- Allow for meaningful comparisons between different datasets or groups.
- Skewness and kurtosis aid in comparing the distribution of variables across different time periods or regions.

Real-World Applications: Shape & Position

MERITSHOT

5. Data Visualization:

- Assist in visualizing data through
  - histograms, box plots, or other graphical representations.
- Visualizations help communicate insights.

MERITSHOT

6. Measures of Standard Scores

Measures of Standard Scores

MERITSHOT

- Also known as z-scores or standard deviations from the mean, are statistical measures.
- Information about how individual data points relate to the mean and standard deviation of a dataset.
- Number of standard deviations a particular data point is away from the mean.

Measures of Standard Scores

- Formula to calculate the z-score for a specific data point:
  - $z = (x - \mu) / \sigma$

Where:

- z is the z-score

- x is the individual data point

- μ is the mean of the dataset

- σ is the standard deviation of the dataset

Key Points: Measures of Standard Scores

1. Interpretation:

- A positive z-score indicates that a data point is above the mean,
    - negative z-score indicates that it is below the mean.
- Magnitude of the z-score represents the distance from the mean in terms of standard deviations.

Key Points: Measures of Standard Scores

MERITSHOT

2. Standardized Comparison:

- By transforming individual data points into z-scores,
  - Different datasets with varying means and standard deviations
  - Can be standardized and compared on a common scale.

Key Points: Measures of Standard Scores

3. Outliers:

- Z-scores can be used to identify outliers.
- Data points with z-scores that fall outside a specific range.
    - (e.g., $z > 3$ or $z < -3$)

Key Points: Measures of Standard Scores

MERITSHOT

4. Percentiles:

- Z-scores can be used to determine the percentile rank of a data point in a distribution.
- Percentile rank corresponds to the area,
    - under the curve of a standard normal distribution.

Key Points: Measures of Standard Scores

MERITSHOT

5. Normal Distribution:

- When a dataset follows a normal distribution,
- 68% of the data falls within one standard deviation of the mean,
- 95% falls within two standard deviations, and
- 99.7% falls within three standard deviations.

MERITSHOT

1. Introduction to Basic Probability

Probability in Real Life Scenarios

MERITSHOT

Weather Forecasting:

- A probability forecast is an assessment.
- An event can occur in terms of percentage and record the risks associated with weather.
- 70% chance that rain may occur.
- Meteorologists use a specific tool and technique to predict the weather forecast.

Probability in Real Life Scenarios

- Other historical databases of the days,
    - Characteristics of temperature, humidity, and pressure, etc.
- 70 out of 100 similar days in the past, it had rained.

Probability in Real Life Scenarios

MERITSHOT

Sport Strategies:

- Understand the strengths and weaknesses of a particular team or player.
- Analysts use probability and odds.
- Regarding the team’s performance and members in the sport.
- Determine in what areas their team is strong.
- In which areas they have to work to attain victory.

Probability in Real Life Scenarios

MERITSHOT

- Gauge the capacity of a particular player in his team.
- Calculate the batting average.
- Batting average in Cricket represents,
    - how many runs a batsman would score before getting out.
- ***For Example:*** If a batsman had scored 30 runs out of 100 from boundaries in the previous match.
- Score 30% of his runs in the next match from boundaries.

Probability in Real Life Scenarios

Politics:

- During the time of election,
  - predict the result of the election.
- Predict the outcome of the election results.

For Example:

- May predict a certain political party to come into power; based on the results of exit polls.

Probability in Real Life Scenarios

MERITSHOT

Insurance:

- Theory of probability or theoretical probability,
    - for framing policy or completing at a premium rate.
- Analyzing the best plan of insurance.

For Example:

- You are an alcoholic, and chances of getting liver disease are higher in you.

Probability in Real Life Scenarios

MERITSHOT

- Go for your health insurance,
  - Chances of you getting sick are higher.

For Example:

- People are getting their mobile phones and laptops insured,
  - Chances of their mobile phones and laptops getting damaged or lost are high.

Probability in Real Life Scenarios

MERITSHOT

Games:

- Getting a desired card when we randomly pick one out of 52.

For Example:

- Probability of picking up a red card in a 52 deck of cards is 26/52;
  - there are 26 red cards in the deck.
- Odds of picking up any other card,

  52/52 – 26/52 = 26/52.

Probability in Real Life Scenarios

MERITSHOT

Lottery Tickets:

- In a Lottery game, each player
  - chooses six distinct numbers from particular range.
- If all the six numbers on a ticket match with that of the winning lottery ticket,
  - ticket holder is a Jackpot winner
  - Order of the numbers.
- Probability of this happening is 1 out of 10 lakhs.

Probability in Real Life Scenarios

MERITSHOT

Natural Disasters:

- Most people take the probability of a natural disaster,
  - when moving into a particular area.
- If a state is hit every year with a major hurricane,

For Example:

- But when an area has never experienced a hurricane before, the probability is much lower.

MERITSHOT

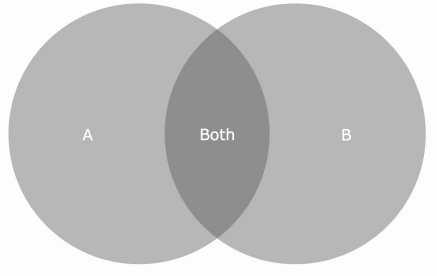
2. Introduction to Set Theory

Relevance in Real World Scenarios

MERITSHOT

1. Venn Diagrams:

- Union, intersection, and difference are visualized using Venn diagrams.
- Use overlapping circles to illustrate the relationships between sets.
- Analyze and understand complex relationships,
    - Market segmentation, customer demographics, or survey responses.



Relevance in Real World Scenarios

2. Probability Calculations:

- Set operations are fundamental to calculating probabilities.
- Help determine the probability of combined events or the probability of multiple conditions occurring simultaneously.
- ***For Example:*** In a survey, the probability of a person being both male and above a certain age can be calculated using set intersection.

Relevance in Real World Scenarios

MERITSHOT

3. Data Analysis and Filtering:

- Set operations are employed in data analysis and filtering tasks.
- ***For Example:*** In data filtering, the intersection of sets can be used to identify records that satisfy multiple conditions simultaneously.
- In data merging, the union of sets can be used to combine datasets with shared elements or merge different categories of data.

Relevance in Real World Scenarios

MERITSHOT

4. Decision-Making:

- Set operations are valuable in decision-making processes.
- Analyze various scenarios, identify common elements or factors, and make informed choices.
- ***For Example:*** In market research, the intersection of customer preferences and demographics can help businesses target specific customer segments.

Sets

1. Union of Sets:

- The union of two sets A and B,
  - A ∪ B,
  - Set that contains all elements that are either in A, in B, or in both.
- Combination of all elements from both sets without duplication.
- Consider outcomes that belong to either of the sets or both.

Sets

2. Intersection of Sets:

- The intersection of two sets A and B,
  - $A \cap B$,
  - Set that contains all elements that are common to both A and B.
- Represents the overlapping elements between the sets.
- Consider outcomes that belong to both sets simultaneously.

Sets

MERITSHOT

3. Difference of Sets:

- The difference of two sets A and B,
  - A - B or A \ B,
  - Set contains all elements that are in A but not in B.
- Represents the elements that are present in the first set
  - But absent in the second set.
- Consider outcomes that belong to one set but not other.

MERITSHOT

3. Introduction to Conditional Probability

Conditional Probability

MERITSHOT

- Probability of the event occurring given another event has already occurred.
- Does not state that there is always a causal relationship between the two events,
  - does not indicate that both events occur.
- Conditional probability of the event A is the probability
  - the event will occur given that an event B has occurred.

Conditional Probability

MERITSHOT

- Conditional probability is written as
  - **P(A|B)**,
- Notation for the probability of A given B.

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

- Here, P(A|B) is the conditional probability i.e.,
  - Probability of event A occurring given that event B has already occurred.

Conditional Probability

MERITSHOT

- P(A∩B) is the joint probability of events A and B i.e.,
  - Probability that both events A and B occur.
- And P(B) is the probability of the event B.
- Applied to the calculation of the conditional probability of events,
  - neither independent nor mutually exclusive.

MERITSHOT

4. Introduction to Bayes Theorem

Conditional Probability

MERITSHOT

- Bayes theorem is named after English statistician.
- Thomas Bayes, who discovered the formula in 1763.
- Foundation of the special statistical inference approach called the Bayes' inference.
- Simple mathematical formula used for calculating conditional probabilities.
- Probability of an event based on prior knowledge of the conditions might be relevant to the event.

Conditional Probability

MERITSHOT

Formula of Bayes Theorem:

- Bayes theorem is a fundamental concept in probability theory and statistics,
    - Allows us to update the probability of a hypothesis based on new evidence.

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

Conditional Probability

MERITSHOT

Where:

- P(A|B) is the probability of event A occurring given that event B has occurred.

- P(B|A) is the probability of event B occurring given that event A has occurred.

- P(A) is the prior probability of event A occurring.

- P(B) is the probability of event B occurring.

Steps in Using Bayes Theorem

MERITSHOT

- Start with the prior probability, P(A),
  - represents your initial belief about the probability of event A occurring.
- Observe new evidence or information, event B.
- Calculate the likelihood, P(B|A),
  - Probability of observing event B given that event A has occurred.

Steps in Using Bayes Theorem

MERITSHOT

- Calculate the marginal probability, P(B),
  - represents the probability of observing event B.
- Use formula to calculate the posterior probability, P(A|B),
  - updated probability of event A occurring given the new evidence.

Steps in Using Bayes Theorem

MERITSHOT

- Bayes theorem is also known,
  - formula for the probability of “causes”.
- Bayes’ theorem is also used in various disciplines,
  - with medicine and pharmacology as the most notable examples.

Example of Bayes Theorem

MERITSHOT

- Find out a patient's probability of having lung disease if they are a smoker. "Being a smoker" is test for lung disease.
- Event A is that the "Patient has lung disease". 10% of patients entering your clinic have lung disease.

  $P(A)=0.1$
- Event B is that the "Patient is a smoker." 5% of the clinic's patients are smokers.

  $P(B)=0.05$

Example of Bayes Theorem

MERITSHOT

- Diagnosed with lung disease, 7% are smokers.
- Probability that a patient is a smoker,
  - they have lung disease is 0.7.
  - P(A)=0.1, P(B)=0.05 and P(B/A)=0.7
  - P(A/B)=0.07 x 0.1/0.05= 0.14
- If the patient is a smoker, their chances of having lung disease is 0.14 (14%).
- Large increase from the 10% suggested by past data.

Example of Bayes Theorem

MERITSHOT

- Let the event A be dangerous fires are rare and the probability of event is 1%,
    - P(A)=0.01
- Event B be smoke is fairly common due to barbecues and the probability of event B is 10%,
    - P(B)=0.1
- 90% of dangerous fires make smoke, it means that the probability of event B given A is 90%,
    - P(B/A)=0.9

Example of Bayes Theorem

- Find the P(A/B).
- We have P(A)=0.01, P(B)=0.1 and P(B/A)=0.9.
- By putting values in the Bayes' theorem,
  - P(A/B)=0.01 x 0.9/ 0.1=0.09
- Probability of a dangerous fire when there is smoke is 9%.

Applications of Bayes Theorem

MERITSHOT

- Bayes' theorem allows scientists to combine a priori beliefs about the probability of an event with empirical evidence,
    - resulting in a new and more robust posterior probability distribution.
- Determine the accuracy of medical test results,
    - how likely any given person is to have a disease and the general accuracy of the test.

Applications of Bayes Theorem

MERITSHOT

- Evaluating the depression tests performance.
- Assess the subsequent probability,
    - person has depression or whether they do not have it.
- Prior probability of information about the diffusion of this pathology,
    - any information of the sensitivity and
    - specificity values of the scores of psychological tests taken by this person.

Applications of Bayes Theorem

MERITSHOT

- Predict the water quality condition.
- Natural and flexible way to approach classification problems.
- Basis of classification problems.
- Used as a building block and starting point for more complex methodologies,
    - the popular Bayesian networks.

MERITSHOT

5. Introduction to Permutations and Combinations

Permutations and Combinations

MERITSHOT

- Permutations and Combinations are very important concepts in Mathematics.
- In which objects from a set may be selected,
  - generally without replacement, to form subsets.
- When the order of selection is a factor,
  - the selection of subsets is called as permutation.
- When the order of selection is not a factor,
  - the selection of subsets is called as combination.

Permutations and Combinations

MERITSHOT

- ab and ba are different arrangements in permutation,
  - same arrangement in case of combinations.
- Main difference between permutation and combination,
  - when the order matters then it is a permutation,
  - when the order doesn’t matter then it is a combination.

Permutations

MERITSHOT

- Collection of objects from a set where the order or the arrangement of the chosen objects does matter.
- Arrangement of objects in a definite order.
- Number of elements should be arranged in a sequence or linear order.
- ***For Example:***
  - Permutation of set A = {3,4} is 2,
  - such as {3,4}, {4,3}.

Permutations

MERITSHOT

- Arrangement of objects in a particular way or order.
- While dealing with permutation,
  - about the selection and arrangement.
- Ordering is very much essential in permutations.
- Permutation is considered as an ordered combination.

Permutations

MERITSHOT

- Represent the permutation as P(n, k),
  - where ‘n’ is the number of objects, and
  - ‘k’ is the number of selections.
- Formula of permutation of ‘n’ objects for ‘k’ selection of objects,

$$P(n,k) = n! / (n - k)!$$

Permutations Example

MERITSHOT

- In how many ways, 2 chairs out of 10 chairs are arranged?
- We have a total of 10 chairs, which means that the number of objects is 10,
  - n = 10 and
  - k = 2.
  - P(10,2) = 10!/(10-2)!
  - 90

Permutations with Repetition

MERITSHOT

- Easiest to calculate.
- Written using the exponent form.
- When the number of objects is “n,” and
  - “k” to be the selection of object,
  - Choosing an object can be in n different ways.
- Permutation of objects when repetition is allowed will be equal to

  =n × n × n × .....(k times).

Permutations with Repetition Example

MERITSHOT

- There are 10 numbers to choose from.
  - Numbers are (0, 1, 2, 3, 4, 5, 6, 7, 8, 9). And we choose 3 of them.
- In how many ways, we can arrange these three numbers from the 10 numbers with repetition?
  - n = 10 and
  - k = 3
  - =10x10x10
  - =1000

Permutations without Repetition

MERITSHOT

- Reduce the number of available choices each time.

$$P(n,k) = n!/(n - k)!$$

- Where,
  - n is the number of things to choose from, and
  - k of them.

Combination

MERITSHOT

- Way of selecting items from a collection where the order of selection does not matter.
- Set of three numbers A, B and C.
- How many ways we can select two numbers from each set.
- “An arrangement of objects where the order in which the objects are selected does not matter.”
- “Selection of things”,
    - where the order of things has no importance.

Formula of Combination

MERITSHOT

- The formula of combination:

$$nCr = n!r!(n - r)!$$

- Where,
  - ‘n’ is the number of things to choose from,
  - choose ‘r’ of them where order doesn’t matter.

Example of Combination

- Find the number of subsets of the set {5, 10, 15, 20, 25, 30, 35, 40, 45, 50} having 3 elements.
  - n = 10
  - r = 3
    - 10xCx3=10!/3!(10-3)!
    - =10x9x8x(7)!/3x2x1x(7)!
    - =120

Example of Combination

MERITSHOT

- In how many ways, we can choose 3 balls from 4 balls?
  - n=4
  - r=3.
    - 4xCx3=4!/3!(4-3)!
    - =4 x 3!/3!
    - =4

Difference Between Permutation & Combination

MERITSHOT

- Permutations are used when order/sequence of arrangement is needed.
- Combinations are used when only the number of possible groups are to be found,
    - order/sequence of arrangements is not needed.

Difference Between Permutation & Combination

MERITSHOT

- Permutation of two things from three given things a, b, c is
  - ab, ba, bc, cb, ac, ca.
- Combination of two things from three given things a, b, c is
  - ab, bc, ca.
    - nPr = n!(n−r)!
- For different possible selection of things,
    - nCr = n!r!(n − r)!

MERITSHOT

6. Introduction to Random Variables

Random Variable

MERITSHOT

- A random variable is a mathematical concept.
- Assigns a numerical value to each possible outcome of a random event or experiment.
- Represents the uncertain outcome of a random process
- Quantify and analyze the probabilities associated with different outcomes.

Random Variable

MERITSHOT

- A random variable can be classified,
  - discrete or continuous,
  - depending on the nature of outcomes.

1. Discrete Random Variable:

- Takes on a countable set of distinct values.
- Be finite or infinite
  - typically integers or whole numbers.

Random Variable

MERITSHOT

- ***Examples of discrete random variables***
- Number of heads obtained when flipping a coin,
- Number of cars passing through an intersection in a given time period,
- or the outcome of rolling a fair six-sided die.

Random Variable

2. Continuous Random Variable:

- Take on any value within a specified range or interval.
- Represents measurements,
    - associated with real numbers.
- **Examples of continuous random variables**
- Height of a person,
- Temperature at a given time, or
- Time taken for a customer to complete a transaction.

Random Variables in Probability for Many Reasons

MERITSHOT

1. Probability Distribution:

- Define and analyze the probability distribution
  - associated with a particular random process.
- Probability distribution describes
  - likelihood of different outcomes
  - provides a basis for calculating probabilities, and
  - performing statistical analyses.

Random Variables in Probability for Many Reasons

MERITSHOT

2. Quantifying Uncertainty:

- A way to quantify uncertainty and randomness in probabilistic events.
- By assigning numerical values to outcomes,
- Random variables allow us to express probabilities, and
- Make predictions about the likelihood of different events occurring.

Random Variables in Probability for Many Reasons

3. Statistical Analysis:

- Random variables are essential for conducting statistical analysis and inference.
- Enable us to model and analyze real-world phenomena,
- Estimate parameters,
- Test hypotheses, and
- Make predictions based on observed data.

Random Variables in Probability for Many Reasons

MERITSHOT

4. Expectation and Variance:

- Calculate important statistical measures,
    - expected value (mean) and variance.
- Insights into the central tendency and variability of the random process under consideration.

Random Variables in Probability for Many Reasons

MERITSHOT

5. Probability Density Functions & Probability Mass Functions:

- Probability density functions (PDFs) for continuous random variables.
- Probability mass functions (PMFs) for discrete random variables.
- Define the probability of different outcomes, and
- Calculate probabilities and perform statistical calculations.

MERITSHOT

7. Introduction to Probability Distribution Functions

Probability Distribution Function

MERITSHOT

- Probability distribution function (PDF) is a mathematical function
    - Describes the probability of each possible outcome of a random variable.
- Way to model the likelihood of different events or values occurring.
- Analyze and understand the behavior of random variables.

Probability Distribution Function

MERITSHOT

- There are two main types of random variables
    - their associated probability distribution functions:

1. Probability Mass Function (PMF):

- PMF is used for discrete random variables.
- Assigns probabilities to each possible value of random variable.

Probability Mass Function

MERITSHOT

- Probability that a random variable takes on a specific value.
- ***The PMF satisfies the following properties:***
  - Probability assigned to each value is non-negative.
  - Sum of probabilities for all possible values is equal to 1.

Binomial Distribution

MERITSHOT

- Describes the number of successful outcomes in a fixed number of independent and identical Bernoulli trials.
- A trial is an individual experiment or observation
  - result in one of two possible outcomes,
  - referred to as "success" or "failure".
- Trials are independent,
  - the outcome of one trial does not affect the outcome of the other trials.

Binomial Distribution

MERITSHOT

- The trials are identical,
  - they have the same probability of success.
- Focuses on counting the number of successes in a fixed number of these trials.

Example

MERITSHOT

- A fair coin and you want to know the probability of getting a certain number of heads when flipping the coin a fixed number of times.
- Each flip of the coin is a trial, and the possible outcomes are "heads" (success) or "tails" (failure).
- Calculate the probability of getting a specific number of heads in a given number of coin flips.

Poisson Distribution

MERITSHOT

- Models the number of events that occur,
    - within a fixed interval of time or space.
- Estimate the likelihood of rare events happening in a given time frame.

Key Characteristics of Poisson Distribution

Events Occur Independently:

- Events being counted are assumed to occur independently of each other.
- Occurrence of one event does not affect the occurrence of another.

Fixed Interval:

- Defined over a fixed interval, such as a specific time period or a designated length.

Key Characteristics of Poisson Distribution

Average Rate:

- Determined by a single parameter called average rate or the average number of events per interval.
    - denoted as λ.

Rare Events:

- Model rare events where the average rate of occurrence is small.

Applications of Poisson distribution Examples

MERITSHOT

- Modeling the number of phone calls received at a customer service center in a given hour.
- Analyzing the number of accidents occurring at a specific intersection per month.
- Estimating the number of emails received per day.
- Calculate the probability of observing a specific number of events within the fixed interval, given the average rate.

Probability Density Function

MERITSHOT

- Continuous random variables.
- PDF does not assign probabilities to individual values.
- Describes the relative likelihood of a random variable falling within a particular range or interval.
- Area under the PDF curve over a given interval
    - represents the probability of random variable falling within that interval.

Probability Density Function

MERITSHOT

- ***PDF satisfies the following properties:***
  - PDF is non-negative for all values within the range.
  - Integral of the PDF over the entire range is equal to 1.

Normal Distribution

MERITSHOT

- Gaussian distribution, is a statistical pattern,
  - describes how values are distributed around a central average or mean.
- Referred to as the "bell curve"
  - due to its characteristic shape.

Normal Distribution

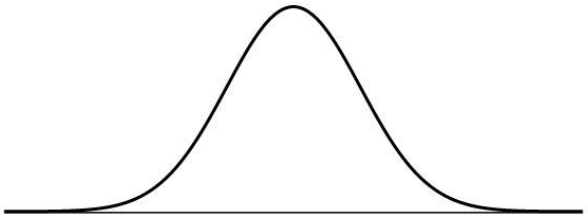
MERITSHOT

- Have a large group of people, and you measure their heights.
- If you were to plot a histogram of their heights, with the tallest people on one end
- Shortest people, you would observe a bell-shaped curve.

Normal Distribution

MERITSHOT

- Have heights around the average or mean height.
- From the average height, the number of people decreases.
- Curve is symmetric,
    - number of people with heights below the average is roughly equal to the number of people with heights above the average.
- By its center (mean), its spread (standard deviation), and its shape (bell-shaped).



Exponential Distribution

MERITSHOT

- Models the time between events in a process,
    - occurs randomly and independently at a constant average rate.
- Events occur continuously and independently over time.
- Imagine you are waiting for a bus at a bus stop.
- The time it takes for the bus to arrive can be modeled using the exponential distribution.

Exponential Distribution

MERITSHOT

- The key idea is that the longer you wait,
  - the more likely it is for the bus to arrive.
- However, the exact arrival time is uncertain and can vary.

Exponential Distribution Characteristics

MERITSHOT

Memorylessness:

- Probability of an event occurring in the next interval of time does not depend on how much time has already passed.

Example:

- If you have already waited for 10 minutes, the probability of the bus arriving in the next minute is the same as if you had just arrived at the bus stop.

Exponential Distribution Characteristics

MERITSHOT

Constant average rate:

- Average rate at which events occur remains constant over time.
- The average waiting time for the bus remains the same,
    - how much time has already passed.

Exponential Distribution Characteristics

MERITSHOT

- Used in various fields,
  - queuing theory, reliability analysis, and finance.
- Understanding and predicting the time it takes for events to happen in systems,
  - events occur randomly and independently at a constant average rate.

Uniform Distribution

MERITSHOT

- All possible outcomes have an equal chance of occurring.
- "Uniform"
  - spread across the range of possible values.
- Imagine you have a bag containing 10 balls, numbered from 1 to 10.
- If you reach into the bag and randomly pick a ball.

Uniform Distribution

MERITSHOT

- Each ball has an equal chance of being selected,
  - distribution of the ball numbers follows a uniform distribution.
- There are no peaks or valleys in the probabilities.
- Every outcome is equally likely,
  - probability of selecting any specific value is the same as selecting any other value

Uniform Distribution

MERITSHOT

- Used when we want to model situations where each outcome has the same likelihood of occurring,
  - rolling a fair die,
  - selecting a random number from a specific range, or
  - randomly selecting a card from a well-shuffled deck.

PDF Essential in Statistics for Many Reasons

MERITSHOT

1. Describing the Likelihood:

- PDFs and PMFs provide a formal way,
  - describe the likelihood of different outcomes
  - or values occurring for a random variable.
- Summarize the probability distribution and allow for analysis and interpretation of the data.

PDF Essential in Statistics for Many Reasons

MERITSHOT

2. Calculating Probabilities:

- Enable the calculation of probabilities associated with specific events or intervals.
- By integrating or summing over the PDF or PMF, probabilities of interest can be computed,
    - probability of a random variable falling within a certain range,
    - or the probability of observing a specific value.

PDF Essential in Statistics for Many Reasons

MERITSHOT

3. Statistical Inference:

- Provide the foundation for estimation and hypothesis testing.
- By comparing observed data to expected distribution based on a probability distribution function.
- Statistical inferences can be made about population parameters or hypotheses.

PDF Essential in Statistics for Many Reasons

MERITSHOT

4. Modeling and Simulation:

- Used to model and simulate random phenomena.
- By selecting an appropriate distribution based on the characteristics of the data,
    - real-world processes can be represented and analyzed.

MERITSHOT

1. Introduction to Normal Distribution

Normal Distribution

MERITSHOT

- Known as the Gaussian distribution or bell curve
  - Continuous probability distribution
    - symmetric around the mean.
- Characterized by its bell-shaped curve,
  - majority of the data points cluster around the mean, with fewer data points in the tails.
- 68-95-99.7 rule,
  - known as the empirical rule or three-sigma rule,
  - Describes behavior of data in a normal distribution.

Normal Distribution

MERITSHOT

- Approximately 68% of the data falls within one standard deviation of the mean.
- About 68% of the data points will fall within interval

$$[\mu - \sigma, \mu + \sigma],$$

  - μ represents the mean and
  - σ represents the standard deviation.
- Approximately 95% of the data falls within two standard deviations of the mean.

Normal Distribution

MERITSHOT

- 95% of the data points will fall within the interval

  $[\mu - 2\sigma, \mu + 2\sigma]$.

- Approximately 99.7% of the data falls within three standard deviations of the mean.
- About 99.7% of the data points will fall within the interval,

  $[\mu - 3\sigma, \mu + 3\sigma]$

Normal Distribution

MERITSHOT

- The 68-95-99.7 rule,
  - quickly assess the spread and variability of data,
  - identify potential outliers, and
  - gain a general understanding of how the data is distributed.
- In exploratory data analysis and helps in making informed decisions,
  - characteristics of a normal distribution.

Normal Distribution Importance & Relevance

MERITSHOT

1. Central Limit Theorem:

- States that the sampling distribution of the mean of a sufficiently large number of independent,
    - identically distributed random variables will be normally distributed.
- Enables us to make inferences about a population based on a smaller sample.

Normal Distribution Importance & Relevance

MERITSHOT

2. Data Analysis:

- Many natural phenomena,
    - heights, weights, IQ scores, errors in measurements,
- Tend to follow a normal distribution.
- Helps in analyzing and modeling such data.

Normal Distribution Importance & Relevance

MERITSHOT

3. Parameter Estimation:

- Assumption of normality is made,
  - allowing for the use of estimation techniques,
  - rely on normality assumptions.
- ***Example:*** In linear regression, maximum likelihood estimation assumes
  - error terms follow a normal distribution.

Normal Distribution Importance & Relevance

MERITSHOT

4. Hypothesis Testing:

- Normal distribution assumptions are often made when performing hypothesis tests,
  - t-tests or analysis of variance (ANOVA).
- Rely on the assumption,
  - data being analyzed are normally distributed.

Normal Distribution Importance & Relevance

MERITSHOT

5. Confidence Intervals:

- In constructing confidence intervals.
- When the population distribution is unknown,
  - but the sample size is sufficiently large,
  - distribution of sample means can be approximated by a normal distribution,
  - allowing for estimation of population parameters.

Normal Distribution Importance & Relevance

MERITSHOT

6. Machine Learning Algorithms:

- Many machine learning algorithms,
  - linear regression,
  - logistic regression, and
  - support vector machines,
- Input features or target variables follow a normal distribution.

Normal Distribution Importance & Relevance

MERITSHOT

- Deviations from normality
- May affect the performance of these algorithms,
    - Help preprocess data appropriately.

Normal Distribution Importance & Relevance

MERITSHOT

7. Outlier Detection:

- Reference for identifying outliers.
- Data points that fall far outside the expected range,
    - number of standard deviations from the mean,
    - can be considered outliers.
- Make accurate assumptions,
- Build appropriate models,
- Perform statistical tests, and
- Interpret results.

Implications of High Standard Deviation

MERITSHOT

1. Increased Dispersion:

- A high standard deviation means
  - data points are, on average, away from mean.
- The values in the dataset are more spread out
  - may be located in a wider range of values.

Implications of High Standard Deviation

MERITSHOT

2. Less Precision:

- A larger standard deviation implies,
  - data points are less precise or
  - have more uncertainty.
- Greater degree of inconsistency or variability among the data values.

Implications of High Standard Deviation

MERITSHOT

3. Higher Risk:

- Higher risk or unpredictability.

Example:

- In financial markets,
  - a high standard deviation in stock prices suggests greater volatility,
  - which can be riskier for investors.

Implications of High Standard Deviation

MERITSHOT

4. Impacted Accuracy:

- If you are using the standard deviation as a measure of the precision or reliability of your data,
  - a high standard deviation can indicate,
  - measurements or observations have more errors or inconsistencies.

Implications of High Standard Deviation

5. Potential Outliers:

- A high standard deviation,
  - presence of outliers in the dataset.
- Outliers are data points,
  - deviate from the rest of the data and
  - can skew the distribution.
- Influential in data analysis and should be examined.

Implications of High Standard Deviation

MERITSHOT

6. Impact on Statistical Inferences:

- High standard deviation may affect the results of statistical tests and estimations.

Example:

- Confidence intervals or hypothesis tests,
  - rely on the assumption of normality,
  - data has a high standard deviation and deviates from normality.

2. Introduction to Skewness and Kurtosis

Skewness and Kurtosis

MERITSHOT

- Skewness and kurtosis are statistical measures,
  - information about the shape and distribution of a dataset.
- Understand the departure of a dataset from a symmetric,
  - bell-shaped distribution, and
  - insights into the presence of outliers, heavy tails, or unusual patterns.

Skewness

MERITSHOT

1. Skewness:

- Skewness measures the asymmetry of a distribution.
- Quantifies the extent to which a dataset's values are concentrated,
    - more on one side of mean compared to other side.

Skewness

MERITSHOT

a) Positive Skewness:

- Also known as right skewness,
- Occurs when the tail of the distribution extends towards higher values.
- Mean is greater than the median.
- ***Example:*** Distribution of household incomes, where a few high-income earners pull the mean income higher, resulting in a longer tail towards the right.

Skewness

MERITSHOT

b) Negative Skewness:

- Also known as left skewness,
- Occurs when the tail of the distribution extends towards lower values.
- Mean is less than the median.
- ***Example:*** Distribution of exam scores, where a few students with very low scores drag the mean down, resulting in a longer tail towards the left.

Skewness

c) Zero Skewness:

- Zero skewness indicates a symmetrical distribution,
    - where the data is evenly distributed around mean.
- The mean and median are equal in this case.

Skewness is Relevant in Data Science

MERITSHOT

Understanding the Distribution:

- Information about the shape of the dataset and indicates if it deviates from a symmetrical distribution.
- Helps in selecting appropriate statistical methods and models for analysis.

Skewness is Relevant in Data Science

MERITSHOT

Identifying Outliers:

- Skewed distributions often indicate the presence of outliers.
- Identifying and addressing outliers,
    - data cleaning and preprocessing tasks.

Skewness is Relevant in Data Science

MERITSHOT

Feature Engineering:

- Skewness can guide feature engineering decisions.
- Transforming skewed variables,
  - help make them more normally distributed,
  - desirable for certain statistical techniques.

Kurtosis

2. Kurtosis:

- Kurtosis measures the tail heaviness or peakedness of a distribution compared to the normal distribution.
- Presence of extreme values or outliers and the concentration of data near the mean.

Kurtosis

a) Mesokurtic:

- A mesokurtic distribution has,
  - kurtosis equal to zero,
  - indicating a similar level of tail,
  - heaviness or lightness as the normal distribution.

Kurtosis

- ***b) Leptokurtic:***
- A leptokurtic distribution has positive kurtosis,
  - heavier tails and
  - higher peak compared to the normal distribution.
- A higher likelihood of extreme values or outliers.
- ***Examples:*** Financial market returns or stock price changes, where extreme events occur more frequently than in a normal distribution.

Kurtosis

c) Platykurtic:

- A platykurtic distribution has negative kurtosis,
  - lighter tails and
  - flatter peak compared to normal distribution.
- Fewer extreme values and less concentration around the mean.
- ***Example:*** Heights of adult humans, which tend to have less variability and are less prone to extreme values.

Kurtosis is Importance in Data Science

MERITSHOT

Identifying deviations from normality:

- Identifying departures from the assumptions of normality.
- Non-normal distributions may require,
    - different statistical techniques or transformations to ensure accurate analysis.

Kurtosis is Importance in Data Science

MERITSHOT

Outlier detection:

- Kurtosis provides insights into the presence of outliers.
- Leptokurtic distributions often indicate,
    - existence of outliers or extreme values.

Kurtosis is Importance in Data Science

MERITSHOT

Risk Assessment:

- Kurtosis is relevant in risk assessment and financial modeling,
    - presence of extreme events or tail risks can impact decision-making.

Kurtosis is Importance in Data Science

Model selection:

- Knowledge of kurtosis helps in choosing appropriate models for analysis.

Example: When fitting linear regression models, kurtosis assumptions may guide the use of robust regression techniques that can handle data with heavy-tailed distributions.

MERITSHOT

3. Introduction to Statistical Transformations

Statistical Transformations Importance

MERITSHOT

1. Normalizing Data:

- Many statistical models and techniques,
  - linear regression,
  - the data follow a normal distribution.
- Transforming the data can help achieve normality,
  - making the models more reliable and accurate.

Statistical Transformations Importance

MERITSHOT

2. Reducing Skewness:

- Impact the validity of statistical inferences and assumptions.
- Transforming the data can reduce skewness,
  - it easier to apply various statistical methods,
  - assume normality.

Statistical Transformations Importance

MERITSHOT

3. Handling Outliers:

- Statistical transformations can mitigate the impact of outliers on data analysis.
- The mean and standard deviation,
    - transformations can reduce their influence and provide a more robust analysis.

Statistical Transformations Importance

MERITSHOT

4. Enhancing Interpretability:

- Transformations can make data more interpretable
  - relationships and making them more linear.
- Aid in understanding the data and improving model interpretability.

Statistical Transformations Importance

MERITSHOT

5. Improving Model Performance:

- Applying appropriate transformations
  - improve the performance of machine learning algorithms.
- Nonlinear transformations can capture complex relationships,
  - while reducing skewness can help models generalize better.

Statistical Transformations Importance

MERITSHOT

6. Overcoming Heteroscedasticity:

- Address heteroscedasticity,
  - variability of the data changes across different levels of the independent variables.
- By stabilizing the variance,
  - transformations can help meet the assumptions of linear regression models.

MERITSHOT

4. Introduction to Sample and Population Mean

What is a Sample?

MERITSHOT

- A sample is just a small part of a whole.

For Example:

- If you work for a polling company and want to know how much people pay for food a year, you aren't going to want to poll over 300 million people.
- Take a fraction of that 300 million;
    - fraction is called a sample.

What is a Sample?

MERITSHOT

- Mean is another word for “average.”
- Sample mean would be the average amount
  - thousand people pay for food a year.
- Allows us to estimate what the whole population is doing, without surveying everyone.
- Let’s say your sample mean for the food example
  - 3500 per year.
- You would get a very similar figure,
  - if you surveyed all 300 million people.

Formula of Sample Mean

MERITSHOT

$$\bar{x} = (\Sigma x_i) / n$$

- x bar is equal to summation of xi divided by n.
- x bar stands for sample mean,
- Summation is the notation which means “add up”,
- xi represents each data point of the sample and
- n is the total number of data points in the sample.

Population Mean

MERITSHOT

- Population mean,
  - denoted by the symbol μ (mu),
- Statistical measure that represents the average value of a variable in an entire population.
- By summing up all the values of the variable in the population.
- Dividing by the total number of individuals in the population.

Population Mean

MERITSHOT

- Measure of the central tendency or average value of the variable of interest.
- Typical or representative value of the variable in question.
- ***For example:*** The variable "income" for a population of individuals.
- Represent the average income of all individuals in the entire population.
- Gain insights into the overall income level.

Population Mean

MERITSHOT

- Formula for calculating the population mean:

$$\mu = (x1 + x2 + x3 + \ldots + xn) / N$$

Where:

- μ is the population mean.
- x1, x2, x3, ..., xn are the individual values of the variable in the population.
- N is the total number of individuals in the population.

Population Mean

MERITSHOT

- Population mean is a parameter,
  - describes the entire population.
- Rely on samples to estimate the population mean.
- The sample mean,
  - $\bar{x}$ (x-bar),
  - used as an estimate of the population mean.

MERITSHOT

5. Introduction to Central Limit Theorem

Central Limit Theorem

MERITSHOT

- The Central Limit Theorem (CLT) is a fundamental concept in inferential statistics.
- Behavior of sample means or sample sums drawn from any distribution.
- Sampling distribution of the mean or sum will approximate a normal distribution,
    - shape of the original population distribution.

Central Limit Theorem

MERITSHOT

- If your sample data has enough data points
  - At least 10% of the population data.
  - Variance then the mean of your sample,
    - equal to the mean of your population data.
- Shape of the original population distribution,
  - as the sample size increases,
  - distribution of the sample means will approximate a normal distribution.

Importance and Examples

MERITSHOT

Example 1: Exam Scores

- Suppose you have a large population of students who have taken an exam.
- Scores on the exam might follow any distribution,
    - skewed distribution or a multimodal distribution.
- ***Central Limit Theorem:*** If you randomly select multiple samples of scores and calculate the mean score of each sample.

Importance and Examples

- The distribution of those sample means will be approximately normal,
  - original distribution.
- Make inferences about the population mean based on the sample means, using techniques like
  - confidence intervals or
  - hypothesis testing.

Importance and Examples

Example 2: Product Quality Control

- Imagine a manufacturing company that produces a large number of products.
- By a continuous variable with a specific distribution.
- Allows you to take multiple random samples from the production line.
- Calculate the average quality of each sample.

Importance and Examples

MERITSHOT

- Distribution of sample means will be
  - Normal,
  - Underlying distribution of individual product quality.
- Estimate the population mean quality,
- Assess the confidence in your estimates, and
- Make decisions about the production process based on the sample means.

Importance in Data Science & Machine Learning

MERITSHOT

Sample Size Independence:

- Central Limit Theorem implies,
  - large enough sample sizes,
  - distribution of sample means tends to be approximately normal,
  - original population is not normally distributed.
- Apply a wide range of statistical techniques,
  - normality.

Importance in Data Science & Machine Learning

MERITSHOT

Hypothesis Testing and Confidence Intervals:

- Use of hypothesis testing and confidence intervals,
  - assumption of normality.
- Statistical inferences and drawing conclusions about population parameters.

Importance in Data Science & Machine Learning

MERITSHOT

Estimation and Prediction:

- Assumption of normality.
- Make accurate estimates and predictions based on the behavior of sample means,
    - original data doesn't follow a normal distribution.

Importance in Data Science & Machine Learning

MERITSHOT

Decision Making:

- Decision-making processes in data science and machine learning.
- Quantify uncertainty,
- Assess the reliability of estimates, and
- Make informed decisions based on statistical evidence.

MERITSHOT

6. Introduction to Bias and Variance

Bias

MERITSHOT

- Systematic error or deviation of an estimator or prediction from the true population value.
- Consistent tendency of the model.
- Method to consistently overestimate or underestimate the true value.
- Bias measures how far off, on average,
    - Predictions or estimates are from the true values.

Bias

MERITSHOT

- A model with high bias is likely to have a
  - significant error, leading to underfitting.
- Underlying relationships in the data,
  - resulting in poor predictive performance, and
  - low flexibility to capture complex patterns.

Variance

MERITSHOT

- Variability or spread of the estimates or predictions over different samples or data points.
- Variance captures how much the estimates or predictions fluctuate or vary,
    - when the model is trained on different subsets of the data.

Variance

- A model with high variance is sensitive to small fluctuations in the training data and tends to overfit.
- Noise or random fluctuations in the data,
    - leading to poor generalization and potentially high errors on unseen data.

Variance

MERITSHOT

Real World Example:

- Consider the task of predicting housing prices based on certain features like square footage, number of bedrooms, and location.
- Bias-variance tradeoff in the context of two types of models:
    - a linear regression model, and
    - a complex polynomial regression model.

Bias

- A model with high bias oversimplifies,
  - underlying relationships in the data and makes strong assumptions.
- A linear regression model assumes a linear relationship between the housing features and the price.
- If the true relationship is more complex,
  - linear model will have a high bias and
  - fail to capture important patterns,
  - resulting in underfitting.

Bias

MERITSHOT

- The linear model consistently,
  - underestimates or overestimates housing prices compared to the true values,
  - leading to a systematic bias.

Variance

- Error introduced due to the model's sensitivity to fluctuations in the training data.
- A model with high variance is sensitive
  - noise or random fluctuations in the training data.
- Polynomial regression model with a high degree.
- Model has a higher flexibility and can fit the training data more.

Variance

MERITSHOT

- Polynomial model captures the noise or random fluctuations in the training data,
  - resulting in high variance.
- It fits the training data extremely well
  - may fail to generalize to unseen data, leading to overfitting.

Bias-Variance Tradeoff

MERITSHOT

- The trade off arises from the inverse relationship between bias and variance.
- Linear regression model has low variance,
  - but high bias,
- Polynomial regression model has low bias,
  - but high variance.
- By adjusting the complexity of the model,
  - navigate the tradeoff.

Bias-Variance Tradeoff

MERITSHOT

For Example:

- Using a polynomial regression model with a moderate degree,
  - strikes a better balance between bias and variance.
- Underlying patterns in the data,
  - while avoiding excessive complexity and overfitting.

Bias-Variance Tradeoff

MERITSHOT

- Finding the right balance between underfitting and overfitting,
  - leading to better predictive performance and generalization to unseen data.
- ***Note:*** Bias-variance tradeoff is not limited to regression models,
  - applies to other machine learning algorithms.

Bias-Variance Tradeoff

MERITSHOT

- Model selection,
- Feature engineering,
- Regularization techniques, and
- Overall iterative process of model improvement.

Applications of Bias and Variance

- Insights into the performance and behavior of statistical models.

Model selection:

- Helps in choosing the appropriate model complexity.
- More complex models tend to have low bias
    - but high variance.

Applications of Bias and Variance

MERITSHOT

- Simpler models have higher bias,
  - but lower variance.
- Depends on the specific problem,
- Available data, and
- Trade-off between underfitting and overfitting.

Applications of Bias and Variance

Performance Evaluation:

- Assess and compare different models.
- Strengths and weaknesses of each model,
- Identify sources of error, and
- Make informed decisions about model improvement.

Applications of Bias and Variance

MERITSHOT

Regularization techniques:

- Reducing variance by introducing a penalty term
  - discourages complex models and
  - promotes simpler ones.
- Regularization can aid in finding a better bias-variance trade-off.

MERITSHOT

7. Introduction to Maximum Likelihood Estimation

Maximum Likelihood Estimation

MERITSHOT

- Determines values for parameters of the model.
- Estimating the parameters of the probability distribution,
    - by maximizing the likelihood function.
- Parameter value that maximizes the likelihood function,
    - maximum likelihood estimate.

Maximum Likelihood Estimation

MERITSHOT

- Principle of Maximum Likelihood Estimation was originally,
  - developed by Ronald Fisher, in the 1920s.
- Probability distribution is the one that makes the observed data “most likely”.
- Parameter vector is considered which,
  - maximizes the likelihood function.

Maximum Likelihood Estimation

MERITSHOT

- In order to implement maximum likelihood estimation,
  - Assume a model, also known as a data generating process, for our data.
  - derive the likelihood function for our data, given our assumed model.
- Once the likelihood function is derived, maximum likelihood estimation
  - more than a simple optimization problem.

Maximum Likelihood Estimation

MERITSHOT

- Make inference about the population,
  - most likely to have generated the sample,
  - the joint probability distribution of the random variables.

Likelihood Function

MERITSHOT

- Fundamental concept in statistical inference.
- Particular population is to produce an observed sample.
- Quantifies the likelihood of observing the given data for different parameter values of a statistical model.
- Observed data are independent and identically distributed.

Likelihood Function

- Joint probability density function (PDF) or probability mass function (PMF) of observed data,
  - treated as a function of the model parameters.
- The likelihood function varies from outcome to outcome of the same experiment,
  - from sample to sample.

Steps to Perform Maximum Likelihood Estimation

- Perform a certain experiment to collect the data.
- Choose a parametric model of the data,
  - modifiable parameters.
- Formulate the likelihood,
  - objective function to be maximized.
- Maximize the objective function and derive the parameters of the model.

Real World Applications

MERITSHOT

- Used in inferential statistics and plays a crucial role in various statistical models,
  - regression models,
  - generalized linear models, and
  - survival analysis.
- MLE provides a principled and efficient method for estimating model parameters,
  - based on observed data.

Real World Applications

MERITSHOT

- Asymptotically unbiased and efficient,
  - converge to the true parameter values as the sample size increases, and
  - minimum variance among consistent estimators.
- Hypothesis testing and constructing confidence intervals,
  - by comparing the likelihoods under different parameter values.

Advantages

- Maximum likelihood estimator is the most efficient estimator.
- Consistent but flexible approach,
  - makes it suitable for a wide variety of applications,
  - cases where assumptions of other models are violated.
- Results in unbiased estimates in larger samples.

Disadvantages

- Assumption of a model and the derivation of the likelihood function.
- Can be sensitive to the choice of starting values.
- Depending on the complexity of the likelihood function,
    - numerical estimation can be computationally expensive.
- Estimates can be biased in small samples.

MERITSHOT

8. Introduction to Confidence Intervals

Confidence Intervals

MERITSHOT

- To estimate population parameters using observed sample data.
- An estimated range of values which is likely to
    - unknown population parameter.
- Estimated range being calculated from a given set of sample data.

Confidence Intervals

MERITSHOT

- Obtained from the observed data that holds the actual value of the unknown parameter.
- Confidence level that quantifies the confidence level,
    - interval estimates the deterministic parameter.
- Based on the Standard Normal Distribution,
- where Z value is the z-score.

Confidence Intervals

MERITSHOT

- A confidence level represents the proportion i.e.,
  - Frequency of acceptable confidence intervals
    - Contain the true value of the unknown parameter.
- Given confidence level from an endless number of independent samples.
- Proportion of the range contains the true value of the parameter.

Confidence Intervals

MERITSHOT

- Confidence level is selected before examining the data.
- Used confidence level is 95% confidence level.
- Confidence levels are also used,
    - 90% and 99% confidence levels.

Confidence Intervals Formula

MERITSHOT

- Used to describe the amount of uncertainty associated with a sample estimate of a population parameter.
- Describes the uncertainty associated with a sampling method.

Confidence Intervals

MERITSHOT

$$\overline{x} \pm z \frac{s}{\sqrt{n}}$$

- Here, x bar is the sample mean,
- z is the confidence level value,
- s is the standard deviation of the sample,
- n is the size of the sample.

- Known as the margin of error.

Example

MERITSHOT

- The table of Confidence Interval and the z-value.
- In this table, the z-values for the confidence interval are given.

Confidence Interval	Z Value
80%	1.282
85%	1.440
90%	1.645
95%	1.960
99%	2.576
99.5%	2.807
99.9%	3.291

Example

MERITSHOT

- In a tree, there are hundreds of apples. You are randomly choosing 46 apples with a mean of 86 and a standard deviation of 6.2. Determine that the apples are big enough. Take the confidence level as 95%.
  - $\overline{x}$ = 86, ∑= 6.2, n = 46
  - 86 ± 1.96 x 6.2/√46.
  - 86±1.79
  - Margin of error = 1.79.
- All the hundreds of apples
  - 84. 21 and 87.79.

MERITSHOT

9. Introduction to Correlations

Correlations

MERITSHOT

- Correlation is a statistical measure,
- Describes the size and direction of a relationship
  - Between two or more variables.
- Does not automatically mean that the change in one variable,
  - Change in the values of the other variable.

Correlations

MERITSHOT

- Measures the strength and direction of a linear relationship,
- For instance:
  - 1 indicates a perfect positive correlation.
  - -1 indicates a perfect negative correlation.
  - 0 indicates that there is no relationship between the different variables.

Correlations

MERITSHOT

Example:

- "Hours worked" and "income earned" would have a positive correlation.
- When the “hours worked” increase,
  - the “income earned” also increases.

Correlations

MERITSHOT

- The two variables "price" and "purchasing power".
- Would have a zero correlation.
- When the Price increases, the purchasing power does not increase.
- ***Example:***
- If we consider the two variables “No. of Leaves”, and “Net Salary”,
  - They would have negative correlation.
- When the No. of Leaves increase,
  - Net salary would decrease.

Causation

MERITSHOT

- Causation indicates that one event is the result of the occurrence of the other event; i.e.,
  - Causal relationship between the two events.
  - Also referred to as cause and effect.
- The relationships between the two types of events are easy to identify.
- An action or an occurrence can cause another.
- ***Example:*** Smoking causes an increase in the risk of developing lung cancer.

Causation

MERITSHOT

- ***Example:*** Where smoking is correlated with alcoholism,
    - but we do not have enough evidence to prove that smoking actually causes alcoholism.
- Smoking and alcoholism are two very different things and may or may not have causation.
- To clearly establish cause and effect, compared with establishing correlation.

Correlation and Causation

MERITSHOT

- To find patterns even when they do not exist.
- When two variables appear to be so closely associated
  - one is dependent on the other.
- Imply a cause and effect relationship
  - where the dependent event is the result of an independent event.

Correlation and Causation

MERITSHOT

- Mostly misunderstood and often used interchangeably.
- Understanding both the statistical terms is very important not only to make conclusions,
  - Making correct conclusions at the end.
- ***Example:*** The study shows that ice cream sales are correlated with homicides in New York.
- As the sales of ice cream rise and fall,
  - the number of homicides.

Correlation and Causation

MERITSHOT

- ***Does the consumption of ice cream cause the death of people?***
- No. Two things are correlated doesn't mean one causes the other.
- Ice cream is not causing the death of people.
- When 2 unrelated things are tied together,
    - either bound by causality or correlation.

Types of Correlation

MERITSHOT

There are 3 major types of correlation:

- Pearson Correlation
- Kendall Rank Correlation
- Spearman Correlation

Pearson Correlation

MERITSHOT

- Used correlation statistic to measure the degree of the relationship between linearly related variables.
- Calculates the effect of change in one variable when the other variable changes.
- ***Example:*** In the stock market, if we want to measure how two stocks are related to each other.
- Measure the degree of relationship between the two.

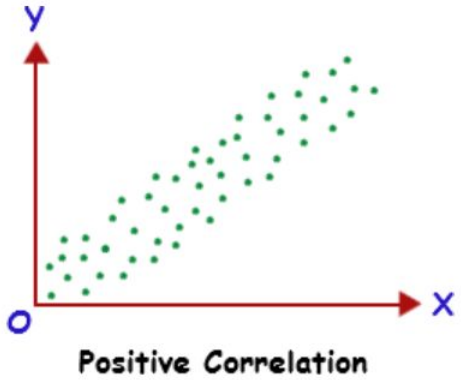
Pearson Correlation

MERITSHOT

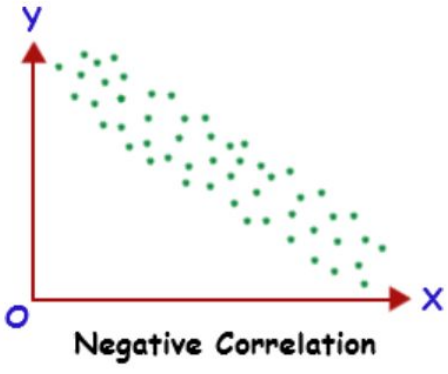
- Value of the Pearson correlation coefficient product is between -1 to +1.
- When the correlation coefficient comes down to zero,
  - the data is said to be not related.
- If we are getting the value of +1,
  - the data are positively correlated
  - -1 has a negative correlation.

Graphs

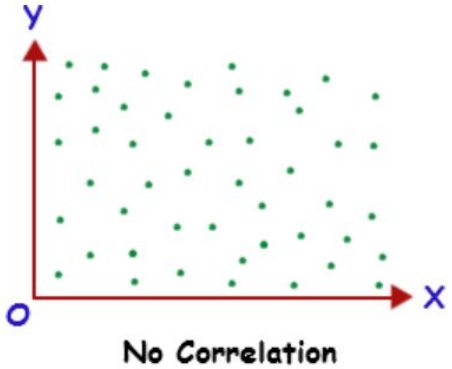
MERITSHOT



Positive Correlation



Negative Correlation



No Correlation

Questions for Pearson Correlation

MERITSHOT

- Is there a statistically significant relationship between age, as measured in years, and height, measured in inches?
- Is there a relationship between temperature, measured in degrees Fahrenheit, and ice cream sales, measured by income?
- Is there a relationship between job satisfaction, as measured by the JSS, and income, measured in dollars?

Assumptions of Pearson Correlation

MERITSHOT

- For the Pearson correlation coefficient r,
- Normally distributed i.e.,
    - have a bell-shaped curve.
- Linearity and homoscedasticity.
- Linearity assumes a straight-line relationship between each of the two variables.
- homoscedasticity assumes that data is equally distributed about the regression line.

Formula of Pearson Correlation Coefficient

MERITSHOT

- For the Pearson correlation coefficient r,
- Normally distributed i.e.,
    - have a bell-shaped curve.
- Linearity and homoscedasticity.
- Linearity assumes a straight-line relationship between each of the two variables.
- homoscedasticity assumes that data is equally distributed about the regression line.

Formula of Pearson Correlation Coefficient

MERITSHOT

- Pearson correlation coefficient is denoted by ‘r’.

$$r = \frac{n\sum xy - \sum x \sum y}{\sqrt{\left[n\sum x^2 - \left(\sum x\right)^2\right]\left[n\sum y^2 - \left(\sum y\right)^2\right]}}$$

- r is the Pearson correlation coefficient,
- x denotes the values in the first set of data,
- y denotes the values in the second set of data
- n is the total number of values.

Kendall Correlation

MERITSHOT

- Kendall correlation,
  - known as Kendall's tau correlation coefficient,
- Measure of association or correlation between two ranked variables.
- Strength and direction of the relationship between variables
  - based on the order or ranking of their values,
  - their specific numerical values.

Kendall Correlation

MERITSHOT

- Kendall correlation can be used when the data is ordinal,
  - Variables being analyzed have a natural order,
  - Intervals between the values may not be equal.

Example:

- Used to analyze rankings, preferences, or ordinal survey responses.

Kendall Correlation

MERITSHOT

Denoted as τ (tau), can range from -1 to 1, where:

- τ = 1 indicates a perfect direct positive correlation,
  - When one variable increases, the other variable also increases consistently.
- τ = -1 indicates a perfect inverse (negative) correlation,
  - When one variable increases, the other variable consistently decreases.
- τ = 0 indicates no correlation or independence between the variables.

Kendall Correlation

MERITSHOT

- Involves comparing pairs of observations and counting the number of concordant and discordant pairs.
- A pair is concordant if the ranks of both variables have the same direction,
    - discordant if the ranks have opposite directions.
- Calculated as the difference between the number of concordant pairs.
- Number of discordant pairs, divided by the total number of pairs.

Kendall Correlation

MERITSHOT

- Robust to outliers and works well with small sample sizes.
- Measure of association for variables,
  - may not have a linear relationship.
- Used in various fields,
  - Social sciences,
  - Economics,
  - Environmental studies,
    - ordinal or ranked data is prevalent.

Kendall Correlation

MERITSHOT

- Represented by the ‘tau’ sign.
- Known as Kendall's tau.
- Non-parametric measure of relationships between columns of ranked data.

Formula of Kendall Rank Correlation

MERITSHOT

- Kendall’s Tau = (C - D)/(C + D)
- C is the number of concordant pairs.
- D is the number of Discordant pairs.

- Concordant means ordered in the same way.
- Discordant means ordered differently.

Example of Kendall Rank Correlation Coefficient

MERITSHOT

- What will be the Kendall Rank Correlation Coefficient for the data having concordant as 61 and discordant as 5?
    - C = 61, D = 5.
- Tau = (C – D)/ (C + D) = (61 - 5)/ (61 + 5) = 56/ 66 = 0.85
- The tau coefficient = 0.85.

Advantages of Kendall's Tau

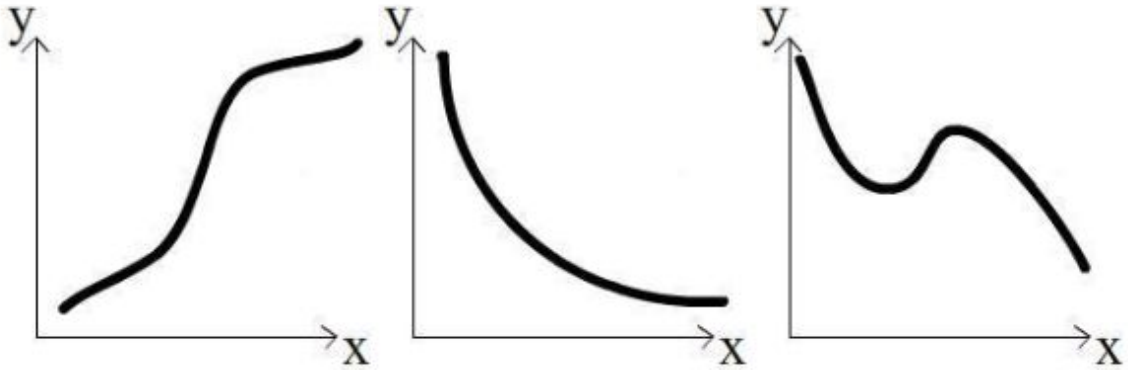
MERITSHOT

- Distribution of Kendall's tau has better statistical properties.
- Probabilities of observing,
  - agreeable (concordant) and
  - non-agreeable (discordant) pairs.
- The interpretations of Kendall's tau and Spearman's rank correlation coefficient,
  - invariably lead to the same inferences.

Monotonic Function

MERITSHOT

- A monotonic function is one that either
  - never increases or never decreases
  - as its independent variable increases.



Monotonically increasing

Monotonically decreasing

Not monotonic

Spearman Correlation

MERITSHOT

- Statistical measure of the strength of a monotonic relationship between paired data.
- In a sample, it is denoted by rs.
- Design constrained as rs is greater than equal to -1 and less than equal to 1.
- Nonparametric measure of rank correlation i.e.,
    - Statistical dependence of ranking between two variables.

Spearman Correlation

MERITSHOT

- Charles Spearman and it is often denoted by
  - Greek letter ‘ρ’ (rho)
  - Used for data analysis.
- Strength and direction of the association between two ranked variables.

Spearman Correlation

MERITSHOT

- A Pearson correlation is a statistical measure of the strength of a linear relationship between paired data.
- The interpretation of Spearman's correlation
  - Similar to that of Pearsons.
- The closure rs is to plus-minus 1,
  - Stronger the monotonic relationship.
- Correlation is an effect size.
- Using the guide for the absolute value of rs.

Spearman Correlation

MERITSHOT

- rs values between 0 and 0.19,
  - Correlation is considered as ‘very weak’,
- rs values between 0.2 and 0.39,
  - Correlation is considered as ‘weak’,
- rs values between 0.4 and 0.59,
  - Correlation is considered as ‘moderate’,
- rs values between 0.6 and 0.79,
  - Correlation is considered as ‘strong’,
- rs values between 0.8 and 1.0,
  - Correlation is considered as ‘very strong’.

Assumptions of Spearman's Correlation

MERITSHOT

- Data must be at least ordinal and the scores on one variable,
    - monotonically related to the other variable.

Note:

- No requirement of normality and hence Pearson's correlation coefficient is a nonparametric statistic.

Formula of Spearman Correlation Coefficient

MERITSHOT

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

- Rho is the Spearman Rank Correlation,
- di is the difference between the ranks of corresponding variables,
- n is the number of observations.

MERITSHOT

10. Introduction to Sampling Methods

Sampling Methods

MERITSHOT

- Sampling is a process used in statistical analysis
    - A predetermined number of observations are taken from a larger population.
- Helps us to make statistical inferences about the population.
- Assume that samples are drawn from the population and sample means and population means are equal.

Sampling Methods

MERITSHOT

- A population can be defined as a whole,
  - includes all items and characteristics of the research taken into study.
- Gathering all this information is time-consuming and costly.
- Make inferences about the population with the help of samples.

Sampling Methods

MERITSHOT

- To sample from a larger population depends on the type of analysis being performed,
    - may include simple random sampling or systematic sampling or even stratified sampling.

Sampling Methods

MERITSHOT

- Sampling technique where every item in population,
  - Chance and likelihood of being selected in the sample.
- Depends on luck or probability.
- Known as a method of chances.
- Easily be a component of a more complex sampling method.
- Every sample has the same probability of being chosen.

Sampling Methods

MERITSHOT

- Random sampling can be applied.
- Method is theoretically simple to understand
  - Difficult to implement.
- Working with a large sample size,
  - Finding a realistic sampling frame.

How Researchers Perform Random Sampling

MERITSHOT

- Prepare a list of all the population members,
  - each member is marked with a specific number.

Example:

- If there are nth members, then they will be numbered from 1 to N.

How Researchers Perform Random Sampling

MERITSHOT

- Researchers choose random samples using two ways:
  - Random number tables
  - Random number generator software.
- Random number generator software,
  - necessary to generate samples.

Example of Random Sampling

MERITSHOT

- We want to select a simple random sample of 100 employees of Company X.
- Assign a number to every employee in the company database from 1 to 1000,
- Random number generator to select 100 numbers.

Advantages of Random Sampling

MERITSHOT

- Random sampling is a fair method of sampling,
  - Helps to reduce any bias involved compared to any other sampling method involved.
- Involves a large sample frame,
  - Easy to pick a smaller sample size from the existing larger population.
- Research doesn’t need to have prior knowledge of the data.

Advantages of Random Sampling

- Gather the researcher need not be a subject expert.
- Fundamental method of collecting the data.
- Require essential listening and recording skills.
- There is no restriction on the sample size that the researcher needs to create.
- From a larger population,
    - you can get a small sample.
- The more samples the better the quality of data.

Systematic Sampling

MERITSHOT

- Researchers use to zero down on the desired population.
- Calculate the sampling interval by dividing the entire population size.
- Extended implementation of probability sampling,
    - each member of the group is selected at regular periods to form a sample.

Systematic Sampling

MERITSHOT

- Probability sampling method where the researcher chooses elements from a target population.
- Random starting point and selects sample members,
  - a fixed ‘sampling interval’.
- Random sampling of a population can be
  - Inefficient and time-consuming,
  - Statisticians turn to other methods.
- When there is a low risk of data manipulation.

Example

MERITSHOT

- To conduct a survey to estimate the average height of students in a school that has 1,000 students.
- You don't have the time or resources to survey every student.
- Use systematic sampling to select a representative sample.

How Systematic Sampling Works

MERITSHOT

1. Define the Sampling Interval:

- Determine the sampling interval.
- By dividing the population size by the desired sample size.

Example:

- If you want a sample size of 100 students, the sampling interval would be 1,000 divided by 100, which gives you a sampling interval of 10.

How Systematic Sampling Works

MERITSHOT

2. Randomly Select a Starting Point:

- To ensure randomness,
    - By randomly selecting a student from the first 10 students.
- Randomly select the 5th student as the starting point.

How Systematic Sampling Works

MERITSHOT

3. Select the Sample:

- Select every 10th student thereafter.
- Select the 5th student, then the 15th student,
    - 25th student.
- Have reached the desired sample size of 100 students.

How Systematic Sampling Works

MERITSHOT

4. Collect Data:

- For each student selected in the systematic sampling process,
    - measure and record their height.
- Used to estimate the average height of all students in the school.

How Systematic Sampling Works

MERITSHOT

- Ensure that every student in the school has an equal chance of being selected.
- Obtaining a representative sample and reduces the potential for bias.
- Assumes that the population,
    - randomly ordered or does not exhibit any systematic patterns.

How Systematic Sampling Works

- Collected the heights of the selected students,
- Calculate the average height of the sample.
- An estimate of the average height of all students in the school.
- Efficient and can provide reliable results when implemented.

Advantages of Systematic Sampling

MERITSHOT

- Simple and convenient for the researchers,
  - create, conduct and analyze samples.
- No need to number each member of a sample,
  - for representing a population in a faster and simpler manner.
- Probability sampling methods,
  - Cluster sampling and stratified sampling.

Advantages of Systematic Sampling

MERITSHOT

- Non-probability methods,
  - Convenience sampling,
- Chances of the clusters created to be highly biased
  - avoided in systematic sampling as the members are at a fixed distance.

Advantages of Systematic Sampling

- Factor of risk involved in this sampling method
  - extremely minimal.
- Diverse members of a population,
  - be beneficial,
  - even distribution of members to form a sample.

Stratified Sampling

MERITSHOT

- Used in statistics and research to divide a population into distinct subgroups or strata.
    - characteristics and then draw samples from each stratum.
- Ensures that the sample is representative of the population's diversity.
- Allows for more accurate estimations and inferences.

Key Steps Involved in Stratified Sampling

MERITSHOT

1. Population Stratification:

- Identify relevant stratification variables or characteristics that divide the population into distinct subgroups.
- Characteristics should be related to research objective.

Example:

- If studying educational outcomes, the stratification variable could be grade level or school type.

Key Steps Involved in Stratified Sampling

2. Stratum Formation:

- The population is divided into mutually exclusive and exhaustive strata.
- Each stratum consists of individuals who share similar characteristics.

Example:

- If the population is students, the strata could be formed by grade level.

Key Steps Involved in Stratified Sampling

MERITSHOT

3. Sample Size Allocation:

- Determine the proportionate or disproportionate allocation of sample sizes,
    - Stratum based on their relative importance or variability.
- Allocation should reflect the population distribution,
    - Ensure adequate representation of each subgroup in the sample.

Key Steps Involved in Stratified Sampling

4. Random Sampling within Strata:

- Random sampling techniques,
    - Simple random sampling or systematic sampling,
- Used to select individuals or units from the stratum.
- Proportional to the stratum's size in the population.

Key Steps Involved in Stratified Sampling

5. Data Collection and Analysis:

- Selected from each stratum,
  - data is collected for the variables of interest.
- Analysis can be performed separately within each stratum or combined,
  - Obtain overall population estimates,
  - Weighting the stratum-specific results.

Example

MERITSHOT

- A company that produces smartphones,
  - Conduct a customer satisfaction survey
  - how satisfied customers are with different models of smartphones.
- Company offers three models:
  - Model A, Model B, and Model C.
- Ensure that your survey sample represents
  - Customer base,
  - Use stratified sampling.

Implement Stratified Sampling

MERITSHOT

1. Population Stratification:

- Divide the population,
  - mutually exclusive and exhaustive groups or
  - Strata based on the smartphone models.
- Three strata:
  - Customers who own Model A,
  - Customers who own Model B, and
  - Customers who own Model C.

Implement Stratified Sampling

MERITSHOT

2. Sample Size Allocation:

- Determine the desired sample size for each stratum,
    - proportion of customers in each group.

Example:

- If 40% of customers own Model A, 30% own Model B, and 30% own Model C,
    - 40% of the total sample for Model A customers,
    - 30% for Model B customers,
    - 30% for Model C customers.

Implement Stratified Sampling

MERITSHOT

3. Random Sampling:

- Randomly select the required number of customers to participate in the survey.

Example:

- If you need 100 participants overall,
  - 40 customers from the Model A group,
  - 30 customers from the Model B group,
  - 30 customers from the Model C group.

Implement Stratified Sampling

MERITSHOT

4. Survey Administration:

- Conduct the customer satisfaction survey on the selected participants.
- Ensure that you collect the relevant feedback and responses.

Implement Stratified Sampling

MERITSHOT

- Ensure that each smartphone model is represented in the sample,
  - its proportion in the population.
- More accurate representation of customer satisfaction for each model,
  - simple random sampling,
  - sample may not adequately capture the distribution of smartphone models.

Implement Stratified Sampling

MERITSHOT

- Helps in achieving a more precise understanding of each subgroup satisfaction levels.
- Enables comparisons across different smartphone models.
- Improves the reliability and accuracy of survey results,
    - valuable insights for decision-making in product development,
    - Marketing, and customer satisfaction improvement efforts.

1. Fundamentals of Hypothesis Testing

Hypothesis

MERITSHOT

- A hypothesis is a guess or assumption,
    - can be tested to see if it is correct or not.

Example:

- Getting at least 6 hours of sleep a day keeps a person fit and enthusiastic for the entire day”.
- “Employees living far from the office usually come late”.

Hypothesis

MERITSHOT

- Many assumptions as we want about certain situations.
- Good hypothesis statement can be tested,
    - Experiments, surveys and other techniques.
- Based on information on prior research.

Types of Hypothesis

MERITSHOT

- ***There are two types of hypothesis:***
  - Null Hypothesis.
  - Alternate Hypothesis.
- A school teacher claims that the average marks scored by all students of the school in Mathematics is greater than equal or to 75.
- Whether the teacher's claim is correct or not.
- Create a Null Hypothesis Statement and an alternate Hypothesis statement.

Types of Hypothesis

MERITSHOT

- Null Hypothesis is the statement,
  - Supports the belief about the population.
- There is no change in the situation.
- Alternate Hypothesis is the statement,
  - Opposes the null hypothesis statement.
- Make in the null hypothesis, the alternate hypothesis is exactly opposite to it.

Types of Hypothesis

MERITSHOT

- Null hypothesis is that there is no change.
  - “The average marks scored by all the students in Mathematics is greater than equal to 75”.
  - Denote Null Hypothesis as H0.
- Alternate hypothesis would be,
  - “The average marks scored by all the students in Mathematics is less than 75”.
  - Denote the Alternate hypothesis by H1.

Types of Hypothesis

MERITSHOT

- Null Hypothesis and Alternate Hypothesis,
    - Test our Hypothesis and either reject the null hypothesis or fail to reject the Null hypothesis.
- “We accept the null hypothesis”.

Types of Hypothesis

- A bulb manufacturer claims that bulbs produced by them have an average lifespan of 500 hours.

Null hypothesis:

- “The average lifespan of the bulb is equal to 500 hours”.

Alternate hypothesis:

- “The average lifespan of the bulb is not equal to 500 hours”.

Types of Errors in Hypothesis Testing

- ***Two types of errors made in hypothesis testing:***
  - Type 1 Error.
  - Type 2 Error.
- Type 1 Error occurs when we reject a True Null hypothesis,
  - Probability of Type 1 Error is denoted by alpha.
- Type 2 Error occurs when we fail to reject a false null hypothesis,
  - Probability of Type 2 error is denoted by beta.

Types of Errors in Hypothesis Testing

MERITSHOT

- “Average marks scored by all the students in Mathematics are greater than equal to 75”.
- Average marks scored by students are pretty less than 75 but the null hypothesis was true.
- Reject a True Null hypothesis.
    - Reject a True Null Hypothesis then it is, Type 1 error.

P-value

MERITSHOT

- P-value is the probability,
  - Null hypothesis is correct.
- P-value stands in support of the null hypothesis.
- Value of P-value lies between 0 to 1.
- If the P-value is less than the level of significance
  - Reject the null hypothesis.

P-value

- If the P-value is greater than the level of significance,
  - We fail to reject the null hypothesis.
- Helps us to determine,
  - whether to reject the null hypothesis or not.

MERITSHOT

2. Introduction to T Tests

Student's T Distribution

MERITSHOT

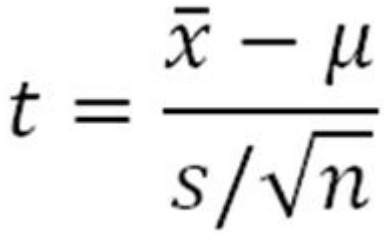
- Student's T distribution which is also known,
  - T distribution is a type of probability distribution
  - Similar to the normal distribution with its bell-shaped curve.
  - Shorter than normal distribution and has heavier tails.
- Less than 30 and the population standard deviation is unknown.

One Sample T-Test

MERITSHOT

- Used to examine or compare the mean of our sample data
  - Known population mean.
- T-test when the sample size is less than 30
  - Population standard deviation is unknown.

One Sample T-Test



$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

- Here,
  - $\bar{x}$ is the sample mean,
  - mu is the population mean,
  - s is the sample standard deviation,
  - n is the sample size.

One Sample T-Test

MERITSHOT

- Using in the T-test is the “degree of freedom”.
- Degrees of freedom is the number of values in the final calculation.
- We calculate the degree of freedom by the simple formula,

  n-1

  - “n” is the sample size.
- A sample size of 20,
  - Degree of freedom = 20-1 = 19.

Example

MERITSHOT

- Assume that the Average Sales Prices for Houses in the US is 180k Dollars.
- Formulate the null hypothesis and alternate hypothesis.

Null Hypothesis:

- No significant difference between the average House Prices in the US.
- Assumption that the House prices in the US is about 180k US Dollars.

Example

MERITSHOT

Alternate Hypothesis:

- There is a statistically significant difference between the average House Prices in the US.
- Assumption that the House prices in the US is about 180k US Dollars.

Two Sample T-Test

MERITSHOT

- Two sample T Test is a statistical procedure used to examine or compare the mean of two separate samples.
- Determine whether there is statistical evidence,
    - Associated population means.

Two Sample T-Test

MERITSHOT

$$t = \frac{\bar{x}_1 - \bar{x}_2}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

- $\bar{x}1$ and $\bar{x}2$ are the sample means.
- n1 and n2 are the sample sizes.
- sp is square is the pooled variance.

$$DF = n_1 + n_2 - 2$$

Paired Sample T Test

MERITSHOT

- Paired Sample T-Test is a statistical procedure for examining or comparing the means of two samples.
- Difference between Two sample t-test and Paired sample t-test,

Two sample t-test:

- Two samples taken were completely different from each other and there was no relationship between the two samples.

Paired Sample T Test

MERITSHOT

Paired samples T-Test:

- Each entity or subject is measured twice.
- Deals with the situation of before and after.

MERITSHOT

3. Introduction to Z Tests

z-test

MERITSHOT

- A z-test is a statistical test.
- Whether two population means are different,
    - when the variances are known.
- Difference between z-test and t-test,
- ***t-test:*** Sample size was less than 30 and the population standard deviation was unknown.
- ***z-test:*** Sample size greater than 30 and the population standard deviation is known.

One Sample z-test

MERITSHOT

- One Sample z-test is used to test whether the mean of a population,
  - Greater than, less than or not equal to a specific value.

$$Z = \frac{\overline{X} - \mu_0}{\sigma / \sqrt{n}}$$

- $\overline{x}$ is the sample mean,
- μ is the population mean,
- σ is the population standard deviation,
- n is the sample size.

z-table

MERITSHOT

- A z-table is a mathematical table.
- Allows us to know the percentage of values below a z-score in a standard normal distribution.
- If the value of z-score is positive,
  - Find the p-value as (1 - z score),
- If the z-score comes out to be negative,
  - z-score is equal to the p-value.

z-table

MERITSHOT

- If the p-value is less than the significance level,
  - Reject the null hypothesis.
- If the p-value is greater than the significance level,
  - Fail to reject the null hypothesis.

Two Sample z-test

$$z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{{\sigma_1}^2}{n_1} + \frac{{\sigma_2}^2}{n_2}}}$$

- $\bar{x}1$ bar and $\bar{x}2$ bar are the means of two samples,
- μ1 and μ2 are the means of population,
- σ1 and σ2 are the population standard deviation,
- n1 and n2 are the sample sizes.

Two Sample z-test

MERITSHOT

Null Hypothesis:

- Means of the first floor and the second floor per feet square houses are equal.

Alternate Hypothesis:

- Means of the first floor and the second floor per feet square houses are not equal.

MERITSHOT

4. Introduction to Chi Squared Tests

Chi Squared Tests

MERITSHOT

- Chi-squared test is used for testing the relationship between categorical variables.
- Null hypothesis of the chi-squared test is that no relationship exists,
    - Categorical variables in the population.
- There are two types of the chi-squared test,
    - Goodness of fit test.
    - Chi-squared test of independence.

Goodness of Fit Test

MERITSHOT

- A goodness-of-fit test is a statistical test.
- Determine whether a sample of data fits,
    - Probability distribution or theoretical model.
- Used in hypothesis testing to assess the adequacy of a model.
- Compare observed data with an expected distribution.

Goodness of Fit Test

MERITSHOT

- Null hypothesis assumes,
  - Observed data follows a specific distribution or model.
- Alternative hypothesis,
  - Observed data deviates significantly from the expected distribution or model.

Goodness of Fit Test

MERITSHOT

- Test involves comparing the observed data with the expected distribution by calculating a test statistic.
- Depends on the specific test being conducted.
- Commonly used goodness-of-fit tests include,
    - Chi-square test,
    - Kolmogorov-Smirnov test,
    - Anderson-Darling test,
    - Cramér-von Mises test.

Relevance & Usefulness of Goodness of Fit Tests

1. Quality Control:

- Goodness-of-fit tests are used to assess.
- Observed data fits the expected distribution in quality control processes.

Example:

- In manufacturing, the tests can determine if a production process follows the expected specifications.

Relevance & Usefulness of Goodness of Fit Tests

MERITSHOT

2. Risk Assessment:

- Goodness-of-fit tests help assess the suitability of statistical models for risk assessment.

Example:

- In finance, the tests can be used to evaluate whether financial data adheres to a particular distribution,
    - the normal distribution.

Relevance & Usefulness of Goodness of Fit Tests

MERITSHOT

3. Epidemiology and Public Health:

- Goodness-of-fit tests are employed to analyze data related to the spread of diseases.
- Assess if the observed data aligns with the expected patterns,
    - Predicted by epidemiological models.

Relevance & Usefulness of Goodness of Fit Tests

MERITSHOT

4. Market Research:

- Aid in determining whether the observed data from surveys.
- Market research studies conform to the expected distributions or models.
- Researchers validate their assumptions and draw accurate conclusions.

Relevance & Usefulness of Goodness of Fit Tests

MERITSHOT

5. Environmental Studies:

- Goodness-of-fit tests are useful for analyzing environmental data.
    - determine if it matches expected patterns.

Example:

- In climate studies, the tests can assess whether temperature or rainfall data adheres to a specific distribution.

Relevance & Usefulness of Goodness of Fit Tests

MERITSHOT

6. Genetics and Genomics:

- Goodness-of-fit tests are applied to genetic data,
    - determine if observed genotypic frequencies match the expected frequencies
    - Predicted by genetic models.

Goodness of Fit Chi-Squared Test

MERITSHOT

$$X^2 = \sum \frac{(o-e)^2}{e}$$

- Chi-squared goodness of fit test is used to determine,
  - Sample data correctly represents the population data or not.

Interpret Chi-Squared Test

MERITSHOT

- Compare it to the critical value or calculate its associated p-value.
- Critical value depends on desired significance level.
  - 0.05
  - Degrees of freedom = -1

Interpret Chi-Squared Test

MERITSHOT

- If the chi-squared test statistic exceeds the critical value,
    - Suggests evidence against the null hypothesis of independence.
- There may be a significant association between the ethnicity categories in the state data.
- Expected frequencies based on the national data.

Interpret Chi-Squared Test

MERITSHOT

- If the chi-squared test statistic is smaller than the critical value,
  - There is no significant evidence to reject the null hypothesis.
- No strong association between the variables and that the observed frequencies are consistent with the expected frequencies based on the national data.

Test of Independence Chi Square test

- Test of independence,
  - Chi-squared test of independence,
- Statistical hypothesis test used to determine,
  - If there is a relationship or association between two categorical variables.
- Assesses whether the observed frequencies of the variables in a contingency table,
- Different from the frequencies,
  - expected if the variables were independent.

Chi-squared Test of Independence Works

MERITSHOT

1. Formulate the Null Hypothesis (H0) and Alternative Hypothesis (H1):

- H0: The two categorical variables are independent,
  - there is no association between them.
- H1: The two categorical variables are dependent
  - there is an association between them.

Chi-squared Test of Independence Works

MERITSHOT

2. Create a Contingency Table:

- A contingency table is a table,
    - Cross-tabulates the observed frequencies of the two categorical variables.

3. Calculate the Expected Frequencies:

- The expected frequencies are calculated,
    - Represent the frequencies that would be expected if the variables were independent.
- Derived from marginal totals of contingency table.

Chi-squared Test of Independence Works

MERITSHOT

4. Calculate the Chi-Square Statistic:

- Chi-square statistic is calculated by comparing the observed.
- Expected frequencies in each cell of the contingency table.
- Measures the discrepancy between the observed and expected frequencies.

Chi-squared Test of Independence Works

MERITSHOT

5. Determine the p-value:

- Chi-square statistic is compared to the chi-square distribution with degrees of freedom calculated based,
  - dimensions of the contingency table.
- p-value is obtained,
  - represents the probability of observing a chi-square statistic,
  - Assuming the null hypothesis is true.

Chi-squared Test of Independence Works

MERITSHOT

6. Make a Decision:

- If the p-value is less than a predetermined significance level,
    - Null hypothesis is rejected,
- Indicating evidence of a relationship or association between the variables.
- If p-value is greater than or equal to significance level,
    - Null hypothesis is not rejected,
- No significant association between the variables.

Relevant & Useful in Real-World Scenarios

MERITSHOT

1. Market Research:

- There is a relationship between variables,
    - Customer demographics and product preferences or purchase behavior.

2. Social Sciences:

- Investigate associations between variables,
    - Educational attainment,
    - Political affiliation,
    - Job satisfaction, and
    - Organizational culture.

Chi-squared Test of Independence Works

MERITSHOT

3. Healthcare:

- An association between a specific treatment.
- Patient outcomes or to examine the relationship,
    - risk factors and disease occurrence.

4. Quality Control:

- Evaluate if there is a relationship between certain factors and product defects.
- Analyze the effectiveness of process changes on the quality of a product.

Chi-squared Test of Independence Works

MERITSHOT

5. Survey Analysis:

- Allows researchers to examine,
    - Whether responses to survey questions are related,
    - Exploring if gender and opinion on a particular issue are associated.

MERITSHOT

5. Introduction to Anova Tests

Anova Tests

MERITSHOT

- Two Sample t-test or Two sample z-test can validate a hypothesis,
    - Containing only two groups at a time.
- Have three or more groups,
    - ANOVA or Analysis of Variance.

Anova Tests

MERITSHOT

- Whether the means of three or more groups are different.
- F-Test to statistically test the equality of means.

$$\text{F-statistics} = \frac{\text{Variation between sample means}}{\text{Variation within sample means}}$$

Anova Tests

MERITSHOT

- ANOVA assumes independence of observations,
- Homogeneity of variances.
- Normally distributed observations within groups.
- Degree of freedom is calculated differently,
    - Numerator and Denominator.
- Number of groups minus 1, and for within groups.
- Total number of observations minus number of groups.