STATISTICS WORKSHEET-1

Q1 to Q9 have only one correct answer. Choose the correct option to answer your question.

1. Bernoulli random variables take (only) the values 1 and 0.
   a) True
   b) False
2. Which of the following theorem states that the distribution of averages of iid variables, properly normalized, becomes that of a standard normal as the sample size increases?
   a) Central Limit Theorem
   b) Central Mean Theorem
   c) Centroid Limit Theorem
   d) All of the mentioned
3. Which of the following is incorrect with respect to use of Poisson distribution?
   a) Modeling event/time data
   b) Modeling bounded count data
   c) Modeling contingency tables
   d) All of the mentioned
4. Point out the correct statement.
   a) The exponent of a normally distributed random variables follows what is called the log- normal distribution
   b) Sums of normally distributed random variables are again normally distributed even if the variables are dependent
   c) The square of a standard normal random variable follows what is called chi-squared distribution
   d) All of the mentioned
5. _____ random variables are used to model rates.
   a) Empirical
   b) Binomial
   c) Poisson
   d) All of the mentioned
6. 10. Usually replacing the standard error by its estimated value does change the CLT.
   a) True
   b) False
7. 1. Which of the following testing is concerned with making decisions using data?
   a) Probability
   b) Hypothesis
   c) Causal
   d) None of the mentioned
8. 4. Normalized data are centered at_____and have units equal to standard deviations of the original data.
   a) 0
   b) 5
   c) 1
   d) 10
9. Which of the following statement is incorrect with respect to outliers?
   a) Outliers can have varying degrees of influence
   b) Outliers can be the result of spurious or real processes
   c) Outliers cannot conform to the regression relationship
   d) None of the mentioned

FLIP ROBO

WORKSHEET

Q10and Q15 are subjective answer type questions, Answer them in your own words briefly.

10. What do you understand by the term Normal Distribution?
11. How do you handle missing data? What imputation techniques do you recommend?
12. What is A/B testing?
13. Is mean imputation of missing data acceptable practice?
14. What is linear regression in statistics?
15. What are the various branches of statistics?

Answer10) Normal distribution, also known as Gaussian distribution, is a probability distribution, that is symmetric about the mean, showing the data near the mean are more frequent in occurrence than data far from the mean. The shape of a normal distribution is often referred to as a “bell curve” because of its characteristic bell-shaped appearance.

Key characteristics of a normal distribution like:

1. Symmetry: The left and right sides of the curve are mirror images of each other.
2. Mean, Median, Mode: In a normal distribution, the mean, median, and mode are equal and located at the center of the distribution.
3. Standard Deviation
4. Denominator of the curve
5. Importance

Answer 11) Handling missing data is a crucial part of data preprocessing in any data analysis or machine learning task. The method used to address missing data often depends on the nature of the data, the mechanism of missingness, and the analysis objectives. Here are some common techniques and recommendations for imputation:

Types of missing data:

1. Missing Completely at Random (MCAR)
2. Missing at Random (MAR)
3. Not Missing at Random (NMAR)

Imputation Techniques:

1. Mean/Median/Mode Imputation:

    - Mean: Replace missing values with the mean of the non- missing values.
    - Median: Use for skewed distributions or ordinal data.
    - Mode: Use for categorical data.
2. K- Nearest Neighbours (KNN) Imputation
3. Multivariate Imputation by Chained Equations (MICE)
4. Regression Imputation
5. Random Forest Imputation
6. Maximum Likelihood
7. Last Observation Carried Forward (LOCF)
8. Interpolation
9. Dropping Missing Values

Recommended:

1. Diagnose Missing Data: Understand the mechanism of missingness and how much data is missing.
2. Choose the Right Method: Match the imputation technique to the type of data and the specific problem you are addressing.
3. Evaluate Impact: After imputing, validate the results by checking impacts on model performance and if possible, compare with baseline models.
4. Report Missingness: Always document how missing data was handled as part of transparency in your analysis.

Answer 12: A/B testing is a method used to compare two versions of a webpage, app, or other user experience to determine which one performs better in achieving a specific goal. The process involves splitting a sample of users into two groups:

1. Group A (Control Group): This group experiences the original version (control) of the content or design.

2. Group B (Variant Group): This group experiences a modified version (variant) with some changes, such as different text, layout, colour schemes, or functionality.

A/B testing is widely used in marketing, wed development, product design, and many other fields to optimize user experiences, increase conversions, and improve overall effectiveness.

Answer 13: Mean imputation is a commonly used method for handling missing data, where missing values for a variable are replaced with the mean of the observed values for that variable. But it has both advantages and disadvantages.

Advantages

1. Simplicity
2. Preserves Sample Size
3. Computational Efficiency

Disadvantages:

1. Bias
2. Underestimation of Variance
3. Distortion of Relationships

While mean imputation can be a quick fix for small amounts of missing data, it is generally not considered the best practice due to its potential to bias results and reduce variability. It's often better to consider other, more robust methods for dealing with missing data that preserve relationships and characteristics of the dataset.

Answer 14: Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables. The primary objective of linear regression is to find the best – fitting linear equation that describes the relationship between the variables.

Key concepts of Linear Regression:

1. Linear Relationship: Linear regression assumes that there is a linear relationship between the dependent variable and the independent variables.
2. Types of Linear Regression:
- Simple Linear Regression:
- Multiple Linear Regression:
3. Method of Least Squares: Linear regression typically uses the method of least squares to estimate the coefficients. This method minimizes the sum of the squared differences between the observed values and the values predicted by the model.

Linear regression is a foundation tool in statistics and data analysis, providing a simple yet powerful way to explore and quantity relationships between variables.

Answer 15: Statistics is a broad field that is generally divided into various branches, each with its own focus, methods, and applications. Here are some of the key branches of statistics:

1. Descriptive Statistics: Focuses on summarizing and describing the features of a dataset.
2. Inferential Statistics: Involves making generalizations or interfaces about a population based on a sample.
3. Probability Theory: The mathematical foundation of statistics that deals with the analysis of random phenomena.

4. Biostatistics: Specialized branch applied to biological and health-related fields.
5. Applied Statistics: Application of statistical methods to real-world problems in various fields such ads economics, engineering, psychology, and social science.
6. Theoretical Statistics: focuses on developing new statistics method and theories.
7. Multivariate Statistics: Deals with the analysis of data involving multiple various simultaneously.
8. Time Series Analysis: Involves analyzing data points collected or recorded at specific time intervals.
9. Non- Parametric Statistics: Methods that do not assume a specific distribution for the data.

Each of these branches can overlap, and they often work together to provide comprehensive insights in research and practical applications across various industries.

15. What are the various branches of statistics?