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
   Answer: a)True

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
   Answer: a)Central Limit Theorem

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

   Answer: b) Modeling bounded count data

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

FLIP ROBO

   Answer: d) All of the mentioned

5. _____random variables are used to model rates.
   a) Empirical
   b) Binomial
   c) Poisson
   d) All of the mentioned
   Answer: c)Poisson

6. Usually replacing the standard error by its estimated value does change the CLT.
   a) True
   b) False
   Answer: b) False

7. Which of the following testing is concerned with making decisions using data?
   a) Probability
   b) Hypothesis
   c) Causal
   d) None of the mentioned

Answer: **b) Hypothesis**

8. Normalized data are centered at_____and have units equal to standard deviations of the original data.
   a) 0
   b) 5
   c) 1
   d) 10

Answer: **a) 0**

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

Answer: **d) None of the mentioned**

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

10. **What do you understand by the term Normal Distribution?**
    - The normal distribution, also known as the Gaussian distribution, is a continuous probability distribution that is symmetric about the mean. It depicts the distribution of a set of data in a bell-shaped curve where most of the observations cluster around the central peak and probabilities for values taper off equally in both directions from the mean. The properties of a normal distribution are defined by its mean and standard deviation.
11. **How do you handle missing data? What imputation techniques do you recommend?**
    - Handling missing data can be approached in several ways depending on the nature and extent of the missingness. Common techniques include:
      - **Deletion Methods**: Removing rows or columns with missing values, useful when the amount of missing data is negligible.
      - **Imputation Methods**: Filling in missing values using various strategies such as:
        - **Mean/Median Imputation**: Replacing missing values with the mean or median of the column.
        - **Mode Imputation**: Replacing with the most frequent value for categorical data.
        - **Predictive Modeling**: Using regression or machine learning models to predict and impute missing values based on other variables.
        - **Multiple Imputation**: Creating multiple imputed datasets and combining results to account for uncertainty in imputations.
      - **Using Algorithms that Handle Missing Data**: Some algorithms like decision trees can handle missing values directly.
12. **What is A/B testing?**
    - A/B testing, or split testing, is a method of comparing two versions of a webpage, app, or other user experience to determine which one performs better. It involves randomly assigning users to one of two groups: the control group (A) that experiences the current version, and the experimental group (B) that experiences the modified version. By analyzing the differences in performance metrics between the two groups, such as click-through rates or conversion rates, one can infer which version is more effective.
13. **Is mean imputation of missing data acceptable practice?**
    - Mean imputation is a simple and commonly used technique, but it has limitations. It can distort the data distribution, underestimate variability, and lead to biased estimates and standard errors. While it can be acceptable for small amounts of missing data or for exploratory analysis, more sophisticated methods like multiple imputation or predictive modeling are generally recommended for more accurate and reliable results.
14. **What is linear regression in statistics?**
    - Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables by fitting a linear equation to the observed data. The simplest form, simple linear regression, involves one independent variable and models the relationship as a straight line. Multiple linear regression involves multiple independent variables and can model more complex relationships. The goal is to estimate the coefficients that minimize the difference between observed and predicted values.
15. **What are the various branches of statistics?**
    - The various branches of statistics include:
      - **Descriptive Statistics**: Summarizing and describing the features of a dataset.
      - **Inferential Statistics**: Making predictions or inferences about a population based on a sample.
      - **Probability Theory**: Studying random phenomena and modeling uncertainty.

- **Mathematical Statistics**: Theoretical foundations and mathematical formulations of statistical methods.
- **Bayesian Statistics**: Incorporating prior knowledge with current evidence to update the probability of a hypothesis.
- **Biostatistics**: Application of statistics in biology and health sciences.
- **Econometrics**: Statistical methods applied to economic data and problems.
- **Psychometrics**: Measuring psychological traits and abilities.
- **Environmental Statistics**: Analyzing and interpreting environmental data.
- **Spatial Statistics**: Analyzing spatial and geographic data.

FLIP ROBO