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 (it takes success and failure represented by 1 and 0)

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 (CLT)

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

Answer: c) The square of a standard normal random variable follows what is called chi-squared distribution

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

Answer: c) Poisson

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

Answer: 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

Answer: b) Hypothesis

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

Answer: a) 0

9. Which of the following statements 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: c) Outliers cannot conform to the regression relationship

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?

Answer: Normal distribution is a fundamental concept in statistics and is commonly used in statistical inference and hypothesis testing. Many natural phenomena, such as heights, weights, and measurement errors, tend to follow a normal distribution. Mean, median, and mode are all equal and located at the center of the distribution. The shape of the distribution looks like a bell.

Normal Distribution, also known as a Gaussian distribution, is a continuous probability distribution that is balanced around its mean.

11. How do you handle missing data? What imputation techniques do you recommend?

Answer: Imputation techniques are methods used to fill in missing values in a dataset. Missing data is a common issue in real-world datasets and can arise for various reasons, such as data entry errors, equipment malfunctions, or survey non-responses. Imputation helps address this issue and allows for more robust analyses.

Handling missing data is an important aspect of data analysis, and the choice of imputation technique depends on the nature of the data and the reasons for the missing values. Analyze the distribution and patterns of missing data to understand their nature and potential impact. Determine whether data is Missing Completely at Random (MCAR), Missing at Random (MAR), or Missing Not at Random (MNAR).

Some of the commonly used imputation techniques are:

Listwise Deletion: Listwise deletion, also known as complete case analysis, entire observations with any missing values are removed from the dataset.

Pairwise Deletion: Pairwise deletion, or available case analysis, involves using all available data for each specific analysis. Instead of excluding an entire observation with missing values, only the specific variables needed for a particular analysis are considered, and missing values in other variables are ignored for that analysis.

Mean Imputation: Replace missing values with the meaning of the observed values for that variable.

Median Imputation: Like mean imputation but uses the median.

Mode Imputation: For categorical data, replace missing values with the mode (most frequently occurring value).

Forward Fill: Forward fill or carry forward replaces missing values with the most recent observance value in the dataset.

Backward Fill: Backward fill or carry backward replaces the missing values with the next observation value in the dataset.

Linear Regression Imputation: Use a linear regression model to predict the missing values based on other variables in the dataset. This method assumes a linear relationship between variables.

K-Nearest Neighbors (KNN) Imputation: Estimate missing values based on the values of their k-nearest neighbors in the dataset.

Multiple Imputation: Generate multiple datasets, each with different imputations for missing values. Analyze each dataset separately and combine the results.

Interpolation: For time series data, interpolate missing values based on the values before and after the missing data points.

Random Forest Imputation: Use a Random Forest algorithm to predict missing values based on other variables in the dataset.

Expectation-Maximization (EM) Algorithm: An iterative algorithm that estimates missing values and imputes them based on the observed data.

Deep Learning Approaches: Neural networks, such as autoencoders, can be used for imputation tasks, especially when dealing with complex relationships in the data.

12. What is A/B testing?

Answer: A/B testing is used to compare two versions of a product or service, it is use to determine which one performs better. The A/B testing is also known as split testing. This method is used in marketing, product development, and other fields. The goal of A/B testing is to identify changes that improve a specific metric or outcome.

A/B testing can be a powerful tool for continuous improvement and optimization in various aspects of business and product strategy.

13. Is mean imputation of missing data acceptable practice?

Answer: Mean imputation is not acceptable in all the situations. It is only acceptable, if the missing data are missing completely at random (MCAR), or missingness is not related to the observed or unobserved data. The variable being imputed is reasonably normally distributed. The imputed variable is not a critical variable in the analysis, and the goal is to preserve the sample size rather than obtaining highly accurate imputations. If Mean imputation does not fit we should use a different model.

14. What is linear regression in statistics?

Answer: Linear regression is probably one of the most important and widely used regression techniques. It's among the simplest regression methods. Linear regression can be defined as a statistical method used to determine the relationship between a dependent variable (also called the response or outcome variable) and one or more independent variables (also called predictors or explanatory variables).

Linear regression can be classified into 2 types based on the number for depend values.
Simple Linear Regression
Multiple Linear Regression

Linear regression is widely used in various fields, including economics, biology, engineering, and social sciences, for modeling and predicting the relationship between variables.

15. What are the various branches of statistics?

Answer: There are many branches of statistics, however, below are the 4 branches that we discussed

1) Descriptive Statistics
2) Inferential Statistics
3) Predictive Statistics /Analytics
4) prescriptive statistics /Analytics