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**

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**

3. Which of the following is incorrect with respect to use of Poisson distribution?
   b) Modeling bounded count data

4. Point out the correct statement.
   c) The square of a standard normal random variable follows what is called chi-squared distribution

5. _____random variables are used to model rates.
   c) Poisson

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?
   b) Hypothesis

8. 4. Normalized data are centered at_____and have units equal to standard deviations of the original data.
   a) **0**

9. Which of the following statement is incorrect with respect to outliers?
   c) Outliers cannot conform to the regression relationship

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-

The normal distribution, also known as the Gaussian distribution, is a fundamental concept in statistics and probability theory. It describes the distribution of a continuous random variable that is symmetric around its mean. When plotted on a graph, data that follows a normal distribution creates a bell-shaped curve, with the peak of the curve at the mean.

One of the defining characteristics of the normal distribution is its symmetry. This means that if you were to fold the curve along its canter, the left and right sides would perfectly match. This symmetry implies that the mean, median, and mode of a normal distribution are all equal and located at the centre of the distribution.

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

Answer-

1. Remove Rows: Delete rows with missing values if the missingness is minimal and won't significantly impact the analysis. However, this may lead to reduced sample size and information loss.
2. Mean/Median/Mode Imputation: Replace missing values with the mean (average), median (middle value), or mode (most frequent value) of the respective column. This is suitable for numerical data.
3. Forward Fill/Backward Fill: Fill missing values with the last known value (forward fill) or the next known value (backward fill). Commonly used for time-series data where values follow a sequence.
4. Predictive Imputation: Use machine learning algorithms like regression or k-nearest neighbors (KNN) to predict missing values based on other variables in the dataset.
5. Multiple Imputation: Generate multiple imputed datasets by replacing missing values multiple times. This method accounts for uncertainty in imputation and provides more robust results.
6. Domain-specific Imputation: Utilize domain knowledge or context-specific information to impute missing values. For example, in medical data, impute based on patient demographics and medical history.

12. **What is A/B testing?**

Answer-

A/B testing is a method used in statistics and data analysis to compare two versions (A and B) of something, typically a webpage, advertisement, or product feature. It involves presenting both versions to different groups of users simultaneously and then analyzing which version performs better based on predefined metrics such as click-through rates, conversion rates, or user engagement. The goal is to determine which version leads to better outcomes and to make data-driven decisions about which variant to implement or continue with.

13. **Is mean imputation of missing data acceptable practice?**

Answer-

Mean imputation of missing data is a commonly used practice, but it has limitations and potential drawbacks. It can distort the original distribution, reduce variability, and introduce bias, especially if the missing data is not missing at random. While it's simple and easy to implement, it's important to consider other imputation methods and evaluate the impact of mean imputation on the validity of your analysis.

14. **What is linear regression in statistics?**
Answer-
Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables It assumes a linear relationship between the variables, where changes in the independent variables are associated with proportional changes in the dependent variable. The goal of linear regression is to find the best-fitting line that minimizes the difference between the observed data points and the predicted values generated by the linear model.

15. **What are the various branches of statistics?**

Answer-
key branches of statistics include:

1. Descriptive Statistics: Summarizing and visualizing data.
2. Inferential Statistics: Making predictions and inferences about populations.
3. Probability Theory: Studying likelihood and random events.
4. Bayesian Statistics: Updating beliefs based on new data.
5. Multivariate Statistics: Analyzing relationships among multiple variables.
6. Time Series Analysis: Studying data collected over time.
7. Experimental Design and A/B Testing: Designing and analyzing experiments.