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.

d) All of the mentioned

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.

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?

Normal distribution, also known as the Gaussian distribution, is a continuous probability distribution that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean. In graph form, normal distribution will appear as a bell curve.

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

There are several ways to handle missing data:
a. Delete rows with missing data
b. Mean/Median/Mode imputation
c. Assigning a unique value
d. Predicting the missing values
e. Using an algorithm which supports missing values, like random forests.
The best method is to delete rows with missing data as it ensures that no bias or variance is added or removed, and ultimately results in a robust and accurate model. However, this is only recommended if there's a lot of data to start with and the percentage of missing values is low.

12. What is A/B testing?

A/B testing is a form of statistical and two-sample hypothesis testing. Statistical hypothesis testing is a method in which a sample dataset is compared against the population data. Two-sample hypothesis testing is a method in determining whether the differences between the two samples are statistically significant or not.

13. Is mean imputation of missing data acceptable practice?

Mean imputation is generally bad practice and can have a negative effect on accuracy when training our ML model. This is because it does not consider any correlations and sometimes generates unrealistic values.

14. What is linear regression in statistics?

Linear regression analysis is used to predict the value of a variable based on the value of another variable. The variable you want to predict is called the dependent variable. The variable you are using to predict the other variable's value is called the independent variable.

15. What are the various branches of statistics?

Statistics have two main branches, namely:

- Descriptive Statistics: This usually summarizes the data from the sample by making use of an index like mean or standard deviation. The methods which are used in the descriptive statistics are displaying, organizing, and describing the data.
- Inferential Statistics: These conclude from data which are subject to random variations like observation mistakes and other sample variation.