STATISTICS WORKSHEET-1

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?
a) Modeling event/time data

4. Point out the correct statement.
b) Sums of normally distributed random variables are again normally distributed even if the variables are dependent

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. 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

Subjective Questions

10. What do you understand by the term Normal Distribution?

Normal distribution, also known as the Gaussian distribution, is a 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. The normal distribution appears as a "bell curve" when graphed.

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

Handling Missing Data:
Missing values are a common issue in machine learning. This occurs when a particular variable lacks data points, resulting in incomplete information and potentially harming the accuracy and dependability of your models. It is essential to address missing values efficiently to ensure strong and impartial results in your machine-learning projects.

1- Mean, Median, and Mode Imputation:

Replace missing values with the mean, median, or mode of the relevant variable.

2. Forward and Backward Fill

Replace missing values with the previous or next non-missing value in the same variable.

3. Interpolation Techniques

Estimate missing values based on surrounding data points using techniques like linear interpolation or spline interpolation.
1 Linear Interpolation
2 Quadratic Interpolation

12. What is A/B testing?
A/B testing is a statistical method used to compare two versions (A and B machine learning model to determine which one performs better on a specific metric

13. Is mean imputation of missing data acceptable practice?
The process of replacing null values in a data collection with the data's mean is known as mean imputation.

Mean imputation is typically considered terrible practice since it ignores feature correlation. Consider the following scenario: we have a table with age and fitness scores, and an eight-year-old has a missing fitness score. If we average the fitness scores of people between the ages of 15 and 80, the eighty-year-old will appear to have a significantly greater fitness level than he actually does.

Second, mean imputation decreases the variance of our data while increasing bias. As a result of the reduced variance, the model is less accurate and the confidence interval is narrower.

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

The two main branches of statistics are descriptive statistics and inferential statistics. Both of these are employed in scientific analysis of data and both are equally important for the student of statistics.