STATISTICS WORKSHEET-1

Question) Bernoulli random variables take (only) the values 1 and 0.

Answer) True

Question) 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 ?

Answer) Central Limit Theorem

Question) Which of the following is incorrect with respect to use of Poisson distribution ?

Answer) Modeling bounded count data

Question) Point out the correct statement.

Answer) 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

Question) _____ random variables are used to model rates.

Answer) Poisson

Question) Usually replacing the standard error by its estimated value does change the CLT.

Answer) False

Question) Which of the following testing is concerned with making decisions using data ?

Answer) Hypothesis

Question) Normalized data are centered at_____and have units equal to standard deviations of the original data.

Answer) 0

Question) Which of the following statement is incorrect with respect to outliers ?

Answer) Outliers cannot conform to the regression relationship

Question) What do you understand by the term Normal Distribution ?

Answer) Normal 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.

Question) How do you handle missing data? What imputation techniques do you recommend ?

Answer) 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.

Question) Is mean imputation of missing data acceptable practice ?

Answer) Mean imputation is generally bad practice because it doesn't take into account feature correlation. For example, imagine we have a table showing age and fitness score and imagine that an eighty-year-old has a missing fitness score. If we took the average fitness score from an age range of 15 to 80, then the eighty-year old will appear to have a much higher fitness score that he actually should.

Question) What is A/B testing ?

Answer) A/B testing is a form of hypothesis testing and two-sample hypothesis testing to compare two versions. The control and variant, of a single variable. It is commonly used to improve and optimize user experience and marketing.

Question) What is linear regression in statistics ?

Answer) Linear regression is one of the statistical techniques used in predictive analysis, in this technique will identify the strength of the impact that the independent variables show on deepened variables.

Question) What are the various branches of statistics ?

Answer) Statistics have two main branches, namely:

a. 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.

b. Inferential Statistics: These conclude from data which are subject to random variations like observation mistakes and other sample variation