STATISTICS WORKSHEET

Q1. Bernoulli random variables take (only) the values 1 and 0.

Ans. a) True

Q2. 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?

Ans. a) Central Limit Theorem

Q3. Which of the following is incorrect with respect to use of Poisson distribution?

Ans. b) Modelling bounded count data

Q4. Point out the correct statement

Ans. d) All of the mentioned

Q5. ______ random variables are used to model rates.

Ans. c) Poisson

Q6. Usually replacing the standard error by its estimated value does change the CLT.

Ans. b) False

Q7. Which of the following testing is concerned with making decisions using data?

Ans. b) Hypothesis

Q8. Normalized data are centred at______and have units equal to standard deviations of the original data

Ans. a) 0

Q9. Which of the following statement is incorrect with respect to outliers?

Ans. c) Outliers can conform to the regression relationship.

Q10. What do you understand by the term Normal Distribution?

Ans. A normal distribution is a symmetric probability distribution about the mean, indicating that data near the mean occur more frequently than data far from the mean. The normal distribution is represented graphically as a "bell curve." Although all symmetrical distributions are normal, not all normal distributions are symmetrical.

Q11. How do you handle missing data? What imputation techniques do you recommend?

Ans. There are many approaches to handle missing data. The most typical response is to disregard it. On the other hand, choosing to make no decision means that your statistical software will decide for you.

Imputation is another tactic used. Imputation involves replacing missing values with an estimate and analysing the complete set of data as if the imputed values were the actual observed values. Some of the Imputation techniques are:

1) Mean imputation

   Determine the mean of all non-missing individuals' observed values for that variable.

   It has the benefit of keeping the mean and sample size constant, but it also has a huge number of disadvantages. Hence this is considered as a bad approach.

2) Substitution

   Assume the value comes from a brand-new individual who wasn't a part of the sample. Or, to put it another way, choose a different topic and use its merits.

3) Imputed regression

   The outcome of modelling the missing variable using other variables to anticipate its value. Instead of using the mean as a result, you are relying on the predicted value, which is influenced by other factors. This preserves the relationships between the variables in the imputation model, but not the range of possible values.

4) Interpolation and Extrapolation

A judgement made on the basis of additional observations made by the same person. Typically, it only functions with data that has been gathered over time. The two types of imputation are single and multiple. Imputation is typically used to refer to a single. The use of only one of the seven approaches to estimate the missing number described above is referred to as "single" estimation. It is well-liked since it is straightforward to comprehend and produces a sample with the same number of observations as the entire data set.

Q12. What is A/B testing?

Ans. A/B testing is a type of experiment in which you divide your website traffic or user base into two groups and provide them with two different iterations of the same web page, app, email, etc. The objective is to compare the results and identify the version that performs better. A/B testing is a type of statistical hypothesis testing or a significance test from the viewpoint of a data scientist.

Q13. Is mean imputation of missing data acceptable practice?

Ans. Mean imputation is frequently seen as a bad approach since it disregards feature correlation. Mean imputation is the process of replacing null values in a data collection with the mean of the data. Mean imputation is frequently seen as a bad approach since it disregards feature correlation. Think about the following situation: we have a table containing age and fitness scores, but the fitness score for an eight-year-old is missing. The elderly person would appear to be far more fit than he actually is if we average the fitness scores of those between the ages of 15 and 80.

Q14. What is linear regression in statistics?

Ans. Linear regression is a basic and often used method of predictive analysis. The two main objectives of regression analysis are to examine: (1) Can a set of predictor variables be used to accurately forecast an outcome (dependent) variable? (2) Which specific variables are highly significant predictors of the outcome variable, as indicated by the size and sign of the beta estimates, and how do they impact the result variable? These regression estimations are used to describe the relationship between a dependent variable and one or more independent variables. The regression equation with one dependent variable and one independent variable is represented by the formula y = bx + c.

Q15. What are the various branches of statistics?

Ans. Descriptive Statistics

CONCEPT - The branch of statistics that focuses on collecting, summarizing, and presenting a set of data.

EXAMPLES - The average age of voters who cast votes for the winning candidate in the most recent presidential election, and the average length of all statistics books.

Inferential Statistics

CONCEPT - The branch of statistics that analyzes sample data to draw conclusions about a population.

EXAMPLE - Suppose it is known what the average grade of 100 students in a certain nation is. Inferential statistics can be used to estimate the country's average student grade using this sample data.