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?
   a)
   b) Modeling bounded count data

4. Point out the correct statement.
   a) All of the mentioned

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

6. 10. Usually replacing the standard error by its estimated value does change the CLT.
   a)
   b) False

7. 1. Which of the following testing is concerned with making decisions using data?
   a)
   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?
   a) 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?

Ans) Normal distribution is the continuous distribution in nature where mean, median and mode all are lined up such that mean will be center. Exactly half lies on the right side and half on the left. It is also called bell curve. Every event is independent of each other in this type of distribution. Most of the values cluster around central region.

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

Ans) Missing data are those values which is null or Nan in the dataset. It shows error when it is sent to machine for analysis and modelling. So we have to use techniques to clear the missing Data.
It is best to avoid Nan values in the dataset as the model will not be 100% accurate with all the imputation techniques we use. If the Nan rows are comparatively very less in proportion to whole data, then I would recommend dropping the data. For eg- in a dataset of 500 rows and less than 10 rows contains Nan then use dropna and drop those rows. But if its more then a Simple Imputer tool can be used.
For string data, simple imputer uses most_frequent value to replace Nan (Uses Mode).
For numeric data, Simple Imputer uses Mean/Median to replace Nan values.

12. What is A/B testing?

Ans) A/B testing is a form of statistical Hypothesis testing or statistical inference. It is used to make decisions that estimates population. We use Null hypothesis and Alternative Hypothesis for the process.

13. Is mean imputation of missing data acceptable practice?

Ans) In my opinion, it really causes bias than they resolve. By taking Mean of the column and replacing with Nan values, it might reduce the error by machine, but for a large data and some number of nan values present, replacing those with mean effects over all model. But if we don't have any other way to get Missing values, then imputing missing data with mean is the best practice and acceptable.

14. What is linear regression in statistics?

Ans) Linear Regression models the relationship between dependent variable (outcome) and independent variable(predictors) by fitting a linear equation Y= a+bx

15. What are the various branches of statistics?

Ans) The two main branches of Statistics are Descriptive Statistics and Inferential statistics.
Descriptive Statistics involves techniques for describing data in a summarized symbolic fashion. It has two further branches Central Tendency and Dispersion of Data. Central Tendency includes Mean, Median and Mode. Dispersion of data includes Range, Variance, Standard Deviation, Skew, Percentile etc.
Inferential Statistics involves drawing inferences from a particular data sample and drawing conclusion using statistics. Z-score values determines the outliers. Hypothesis testing and Estimation testing are the two methods in Inferential Statistics. Hypothesis testing includes T-test, ANOVA test, Chi square test, F-test, etc

FLIP ROBO