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.

Answer: 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?

Answer: a) Central Limit Theorem

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

Answer: b) Modeling bounded count data

4. Point out the correct statement.

Answer: d) All of the mentioned

5. ______ random variables are used to model rates.

Answer: c) Poisson

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

Answer: b) False

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

Answer: b) Hypothesis

8. Normalized data are centered at______and have units equal to standard deviations of the original data.

Answer: a) 0

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

Answer: 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?

Answer: Normal distribution is probability bell curve. Normal distribution has two parameters - mean and standard deviation. In Normal distribution the mean is zero and standard deviation is 1. Normal deviation is important in statistics and is key to the CLT.

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

Answer: To handle a missing data we need to detect missing values numerically and visually using Missingno library. After classified the patterns in missing values, it needs to treat them by deletion technique and then imputation technique.

The imputation technique replaces missing values with substituted values. The missing values can be imputed in many ways depending upon the nature of the data and its problem. Imputation techniques can be broadly they can be classified as follows:

Imputation with constant value: It replaces the missing values with either zero or any constant value.

Imputation using Statistics: The syntax is the same as imputation with constant only the SimpleImputer strategy will change.It can be “Mean” or “Median” or “Most_Frequent”.

K_Nearest Neighbor Imputation: The KNN algorithm helps to impute missing data by finding the closest neighbors using the Euclidean distance metric to the observation with missing data and imputing them based on the non-missing values in the neighbors.

12. What is A/B testing?

Answer: A/B testing, also known as split testing, refers to a randomized experimentation process wherein two or more versions of a variable (web page, page element, etc.) are shown to different segments of website visitors at the same time to determine which version leaves the maximum impact and drives business metrics.

13. Is mean imputation of missing data acceptable practice?

Answer: 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. 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?

Answer: Linear regression models the relationships between at least one explanatory variable and an outcome variable. These variables are known as the independent and dependent variables, respectively. When there is one independent variable (IV), the procedure is known as simple linear regression. When there are more independent variable (IV), it is known as multiple regression. Linear regression has two primary purposes–understanding the relationships between variables and forecasting.

15. What are the various branches of statistics?

Answer: Descriptive statistics and inferential statistics are the two main branches of statistics. Both of these are used in scientific data analysis and are equally significant. Descriptive statistics deals with the presentation and collecting of data. Inference statistics are statistical techniques that allow to utilise data from a sample to conclude, predict the behaviour and make judgments or decisions.