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?

b) Modeling bounded count data

4. Point out the correct statement.

d) All of the mentioned

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. 1. 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

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 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 is the proper term for a probability bell curve.

In a normal distribution the mean is zero and the standard deviation is 1. It has zero skew and a kurtosis of 3.

Normal distributions are symmetrical, but not all symmetrical distributions are normal.

Many naturally-occurring phenomena tend to approximate the normal distribution.

In finance, most pricing distributions are not, however, perfectly normal.

0 seconds of 0 seconds

The normal distribution is the most common type of distribution assumed in technical stock market analysis and in other types of statistical analyses. The standard normal distribution has two parameters: the mean and the standard deviation.

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

Ans. **Missing data can saved from mishandling in the following ways:**

Missing data reduces the statistical power of the analysis, which can distort the validity of the results,

One way of handling missing values is the deletion of the rows or columns having null values.

If any columns have more than half of the values as null then you can drop the entire column.

In the same way, rows can also be dropped if having one or more columns values as null.

These are the following Imputation techniques:

1. Complete Case Analysis(CCA):-

This is a quite straightforward method of handling the Missing Data, which directly removes the rows that have missing data i.e we consider only those rows where we have complete data i.e data is not missing.

This method is also popularly known as “Listwise deletion”.

2. Arbitrary Value Imputation

This is an important technique used in Imputation as it can handle both the Numerical and Categorical variables.

This technique states that we group the missing values in a column and assign them to a new value that is far away from the range of that column.

Mostly we use values like 99999999 or -9999999 or “Missing” or “Not defined” for numerical & categorical variables.

3. Frequent Category Imputation

This technique says to replace the missing value with the variable with the highest frequency or in simple words replacing the values with the Mode of that column.

This technique is also referred to as Mode Imputation.

12. What is A/B testing?

Ans. 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.

Essentially, A/B testing eliminates all the guesswork out of website optimization and enables experience optimizers to make data-backed decisions.

In A/B testing, A refers to ‘control’ or the original testing variable.

Whereas B refers to ‘variation’ or a new version of the original testing variable.

13. Is mean imputation of missing data acceptable practice?

Ans. 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?

Ans. 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.

Linear regression is a basic and commonly used type of predictive analysis.

This form of analysis estimates the coefficients of the linear equation, involving one or more independent variables that best predict the value of the dependent variable.

Linear regression fits a straight line or surface that minimizes the discrepancies between predicted and actual output values.

15. What are the various branches of statistics?

Ans. These are the following branches od statistics:

1) Descriptive Statistics

Descriptive statistics deals with the collection of data, its presentation in various forms, such as tables, graphs and diagrams and finding averages and other measures which would describe the data.

For example: Industrial statistics, population statistics, trade statistics, etc. Businessmen make use of descriptive statistics in presenting their annual reports, final accounts, and bank statements.

(2) Inferential Statistics

Inferential statistics deals with techniques used for the analysis of data, making estimates and drawing conclusions from limited information obtained through sampling and testing the reliability of the estimates.

For example: Suppose we want to have an idea about the percentage of the illiterate population of our country. We take a sample from the population and find the proportion of illiterate individuals in the sample. With the help of probability, this sample proportion enables us to make some inferences about the population proportion. This study belongs to inferential statistics.