STATISTICS WORKSHEET-1

Q.1Bernoulli random variables take (only) the values 1 and 0.
A- a) True

Q.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-a) Central Limit Theorem

Q.3 Which of the following is incorrect with respect to use of Poisson distribution?
b) Modeling bounded count data

Q.4 Point out the correct statement.

Q.5 _____ random variables are used to model rates.
c) Poisson

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

Q.7 Which of the following testing is concerned with making decisions using data?
c) Hypothesis

Q.8 Normalized data are centered at_____and have units equal to standard deviations of the original data
a) 0

Q. 9 Which of the following statement is incorrect with respect to outliers?
C)Outliers cannot conform to the regression relationship

Q.10 What do you understand by the term Normal Distribution?
Normal distribution, 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. In graph form, normal distribution will appear as a bell curve. 1:13. 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.

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

Firstly, detect the quantity of missing value in every column of dataset, it will give us idea about the distribution of missing values. Using heatmap we find out the pattern of missing value. After classifying the pattern in missing value, we need to treat them.

Imputation Techniques:

The imputation techniques replace missing values with substituted values. The missing values can be imputed in many ways depending upon the nature of the dataset and its problem. Imputed techniques can be broadly classified as:

Imputation with constant value:

As its name suggests, it replaces the missing values with either zero or any constant value. We use simple imputer class from sklearn for this.

Imputation using statistics:

In this also we use simple imputer class from sklearn however the strategy is changed. It can be 'mean' ,'median', 'Most Frequent'.

Mean will replace missing values using the mean in each column. It is preferable if data is numeric and non skewed.

Q.12 What is A/B testing?

A/B testing is also known as split testing, refers to a randomized experiment process where two or more versions of a variable are shown to different segments at same time to determine which version leaves the maximum impact and drives business metrics.

A/B testing eliminates all the guesswork out and enable experience optimizers to make data backed decisions. Here A refers to 'control' or the original testing variable whereas B refers to a 'variation' or a new version of the original testing variable.

Q.13 Is mean imputation of missing data acceptable practice?

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, an eight-year-old has a missing fitness score. If we average the fitness the fitness scores of people between ages of 15 and 80, the eighty-year-old will appear to have a significantly grater fitness level than he actually does.

Second, mean imputation decreases the variance of our data while increasing bias. As a result of reduced variance, the model is less accurate and the confidence interval is a narrower.

Q.14 What is linear regression in statistics?

In statistics, linear regression is a linear approach for modeling the relationship between scalar response and one or more explanatory variables. Linear regression has two primary purposes- understanding the relationships between variables and forecasting.

In one dimensional form it is represented as y=mx+c, where m is the coefficient and C is the intercept. The coefficients represent the estimated magnitude and direction(positive/negative) of the relationship between each independent variables and dependent variable.

A linear regression equation allows you to predict the mean valve of the dependent variable that you specify.

Q15 What are the various branches of statistics?

Statistics is the branch of mathematics that deals with data. Data is a collection of values. A collection of data is often referred to as a data set or set of data, but other words such as a list or simply collection are also often used.

These are the branches of statistics:

1. **Descriptive Statistics:**
   It deals with the presentation and collection of data. This is usually the first part of statistics analysis.
2. **Inferential Statistics:**
   Descriptive statistics forms the basis for analysis and discussion in such diverse fields.