STATISTICS WORKSHEET 1

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

ANS : - (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 ?

ANS : - (A) Central Limit Theorem.

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

ANS : - (B) Modeling bounded count data.

4 > Point out the correct statement.

ANS : - (D) All of the Mentioned.

5 > _______ random variables are used to model rates

ANS : - (C) Poisson

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

ANS :- (B) False.

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

ANS :- (B) Hypothesis.

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

ANS : - (A) 0

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

ANS :- (C) Outliers cannot conform to the regression relationship.

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. In graph form, normal distribution will appear as a bell curve.

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

ANS : - Use deletion methods to eliminate missing data

The deletion methods only work for certain datasets where participants have missing fields. The problem with this method is that it may not be practical for large datasets

Use regression analysis to systematically eliminate data
Regression is useful for handling missing data because it can be used to predict the null value

using other information from the dataset. There are several methods of regression analysis, like Stochastic regression. Regression methods can be successful in finding the missing data, but this largely depends on how well connected the remaining data is. Of course, the one drawback with regression analysis is that it requires significant computing power, which could be a problem if data scientists are dealing with a large dataset.

Data scientists can use data imputation technique

Data scientists use two data imputation techniques to handle missing data: Average imputation and common-point imputation. Average imputation uses the average value of the responses from other data entries to fill out missing values. However, a word of caution when using this method - it can artificially reduce the variability of the dataset. Common-point imputation, on the other hand, is when the data scientists utilise the middle point or the most commonly chosen value

12 > What is A/B testing?

ANS : - A /B testing

, also known as split testing, refers to a randomized experimentation
Answer.
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 drive 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.

15 > What are the various branches of statistics?

ANS : - There are two main branches of statistics

- Inferential Statistic.
- Descriptive Statistic.

Inferential Statistics:

Inferential statistics used to make inference and describe about the population. These stats are more useful when its not easy or possible to examine each member of the population

Descriptive Statistics:

Descriptive statistics are use to get a brief summary of data. You can have the summary of data in numerical or graphical form.

There are two types of Descriptive Statistics:

1.Central Tendency

a. Mean
b. Median
c. Mode

2. Dispersion of Data

a. Range
b. Variance
c. Standard Deviation
d. Skewness/Percentile.