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
   b)

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
   b)
   c)
   d)
3. Which of the following is incorrect with respect to use of Poisson distribution?
   a)
   b)
   c)
   d) All of the mentioned

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

5. _______random variables are used to model rates.
   a)
   b)
   c)
   d) Poisson
   e) All the of above. Poisson will give better result.

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)
   b)
   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 data symmetrically distributed which are equally spread on the plane.

It forms a bell-shaped curve, with maximum data points in the central part of the curve, which is equal on both sides.

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

Ans= Handling missing data can be handled by following methods
1. Filling the missing values -Nan or Null in empty or blank spaces.
2.Deleting the missing values- Dropping the columns in the dataset. Through dropna() , but through the drop method if maximum data is drop then, we won't get the good accuracy.
3. Imputation the missing values by various methods

1. Through Mean- Finding average and fill by average data.
Median – the middle value and filling the data. (average of two terms between the missing value)
Mode-Used by treating values with the maximum repeated values in the dataset.

2.- KNN – Finding the nearest neighbor and allotting to the set of datapoints and used in treating many scatter data values.
3 Model selection like Simple Imputer which is avaible in Sk.learn library.

As per the imputation techniques mostly commonly used is fillna () , Dropna() and through mean, median and mode.

12. What is A/B testing?
Ans= A/B testing or split testing, it is to compare two iterations and the performance differences between them to choose the best among them.

we need to choose two variants of the same content, with minimum variation. These two variants or two groups with identical result and difference or compare their performance over a certain period of time to get the best results out of both the variant this is known as A/B testing.

13. Is mean imputation of missing data acceptable practice?
Ans= Yes, always, as Mean is the quick, common and easy approach, but it can introduce bias if the missing data
is not randomly distributed. Treating missing values of numeric column but if there are outliers,
then the mean will not be correct method as outliers need to be treated first
.and then use 'fillna ()' method for imputing.

14. What is linear regression in statistics?
Linear regression is relationship or prediction between dependent variable and one or more independent variable with a linear equation. Predicts the unknown data based on known data with an equation y=mx+c.

there are 2 types simple linear regression and multiple linear regression

y=mx +c
y=m1x1+m2x2+……..mnxn+c
X is independent variable.
Y is dependent variable
M is slope
C is intercept

15. What are the various branches of statistics?

The Statistics is collecting and analyzing numerical data in large quantities.
There are two types of Statistics and they are
1. Descriptive Statistics
2.Inferential Statistics

Descriptive Statistics gives a description of the data either through numerical calculated graphs or tables.

It is simply used for summarizing data. There are two categories in this as follows.

Measure of Central Tendency
Measure of Variability

Inferential Statistics makes inferences and predictions based on a sample of data taken. It generalizes a large dataset and applies probabilities to draw a conclusion.

Inferential Statistics is mainly related to hypothesis testing and to reject the null hypothesis.