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.

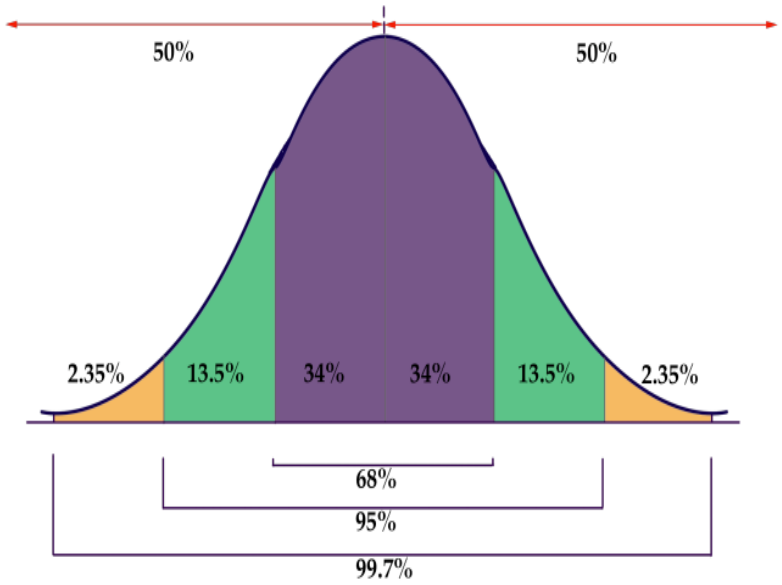
10. What do you understand by the term Normal Distribution?

A. Normal Distribution is a probability distribution of continuous data, where in data is distributed symmetrically

Most of the data points are close to the Mean

It is also known as the bell curve

In a Normal Distribution Mean=Median= Mode



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

A. Missing values (NaN or Null) are common in data samples. It may be due to errors on the data collection, blank spaces on surveys, measurements not applicable…etc

They need to be taken care of as they

Reduce the efficiency of the ML model.

Affect the overall distribution of data values.

and lead to a biased effect in the estimation of the ML model

Eliminating the samples or features with missing values we risk to delete relevant information or too many samples

Hence we use imputation techniques to deal with them

Imputation can be done using any of the below techniques–

Impute by mean

Impute by median

Knn Imputation etc.

12. What is A/B testing?

A. A/B testing also known as split testing is used to compare two variants of same sample

In Data Science we build different models for the same dataset and check which gives us the best result

We also use techniques like Hyperparameter tuning to get better results

13. Is mean imputation of missing data acceptable practice?

A. Imputation of data using mean can be accepted for continuous data

However its is not advisable or cannot be used for categorical data.

14. What is linear regression in statistics?

Linear Regression is one of the most fundamental and widely known Machine Algorithms

It is used on discreet or continues data

Linear Regression predicts the dependent variable (label) using a regression line based on the independent variables (features)

The line of regression shows the relationship between the label and the features

Equation: Y= mx+c

15. What are the various branches of statistics

A. There are two main branches of statistics:

1. Descriptive Statistics: Deals with summarizing and presenting data in a readable ,easily understood form

Eg: Graphs, Tables ,Charts etc.

2. Using data collected from a small group (Sample) to predict an outcome.

Eg. Hypothesis Testing