FLIP ROBO

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) Modelling 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

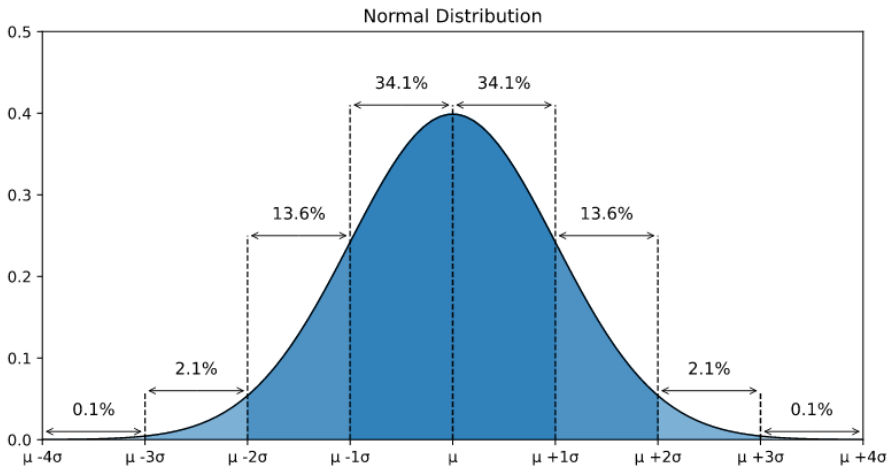
Q10and Q15 are subjective answer type questions, Answer them in your own words briefly.

10. What do you understand by the term Normal Distribution?
The normal distribution is an important probability distribution used in statistics it is often referred to as a 'bell curve' because of its shape.

- Most of the values are around the centre (μ)

- The median and mean are equal
- It has only one mode
- It is symmetric, meaning it decreases the same amount on the left and the right of the centre

  The area under the whole curve is equal to 1, or 100%



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

We can handle missing data by deleting Rows with missing values. By Impute missing values for continuous variable and impute missing values for categorical variable. KNN imputer, Iterative Imputers are also used to handle the missing values.

There are many techniques that we are using in imputation, I can't recommend anyone technique but depending upon the situation like when missing data is numerical, we use mean or median values if data is categorical, we use mode value sometimes we impute based on class sometimes we impute based on model also. Multiple imputation is advantageous because it provides within and between imputation variability.

12. What is A/B testing?

A/B testing is statistical hypothesis testing, it is a process whereby a hypothesis is made about the relationship between two data sets and those data sets are then compared against each other to determine if there is a statistically significant relationship or not

13. Is mean imputation of missing data acceptable practice?

Not always, because outliers' data point will have significant impact on the mean hence it is not recommended to use mean for replacing the missing value. Replacing missing value with mean does not create great model sometimes result will be biased.

14. What is linear regression in statistics?

Linear Regression analysis is predicting the value of one variable based on the value of another variable. The variable we want to predict is dependent variable and the variable we are using to predict the other variable's value is called independent variable.

15. What are the various branches of statistics

Various branches of statistics are

a) Descriptive Statistics and
b) Inferential Statistics