STATISTICS WORKSHEET-1

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

Ans. b) False

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.

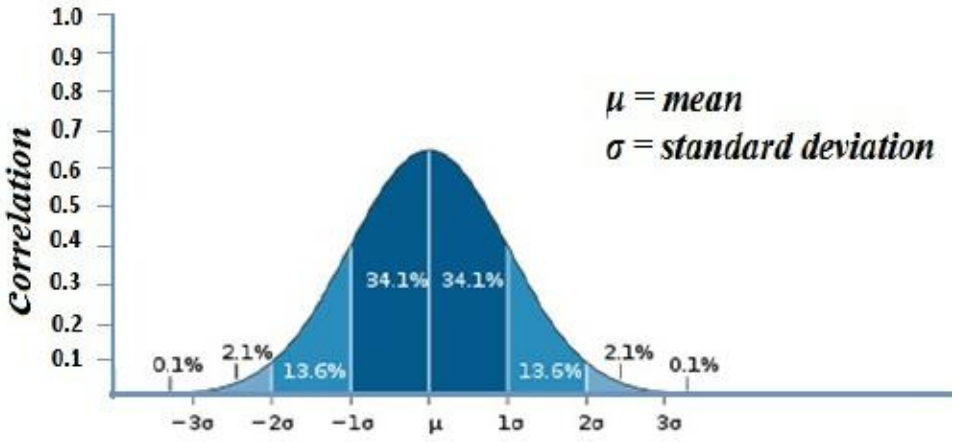
Ans. a) 0

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, also known as Gaussian Distribution is perhaps the most significant and continuous probability distribution. It can also be referred to as a bell curve where a huge number of random variables are either nearly or exactly represented by the distribution. In this distribution the mean is zero (0) and the standard deviation is 1.



The above graph is a representation of a Normal distribution.

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

Ans. There are mainly 2 primary methods to deal with missing data, viz., imputation and the removal of data.

Through imputation method, missing data is filled on the basis of reasonable guesses. This can be helpful when the percentage of missing data is significantly less. However, if the missing percentage is high this method can impact the model in a negative way.

On the other hand, removing or deleting data might not be the best way to predict a model because it may lead to unreliable analysis. Infact there are chances or missing out on key observations of specific factors of the model.

In my view Arbitrary Value Imputation should be chosen while dealing with missing data because it is easy to implement, can be used in production and helps to retain the importance of missing values. However, one needs to be extra cautious while selecting an arbitrary value.

12. What is A/B testing?

Ans. A/B tests, also known as split tests, allows one to compare two versions of something to learn and find out which one is more effective.

13. Is mean imputation of missing data acceptable practice?

Ans. Mean imputation is acceptable only when the missing value proportion is not large enough.

14. What is linear regression in statistics?

Ans. Linear regression is a commonly used predictive analysis that helps to analyze whether a set of predictor variables predicts an outcome (dependent) variable correctly and which of the variables in particular are significant predictors for the outcome.

The formula y = mx + c, represents the relationship between one dependent variable (y) and one independent variable(x), where c is constant, m is the regression coefficient.

15. What are the various branches of statistics?

Ans. There are two main branches of statistics, viz., descriptive statistics and inferential statistics.

Descriptive Statistics - It deals with the presentation and collection of data.

Inferential Statistics – It helps to reach to the right outcome after thorough analysis of the statistical data.