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) False

Answer: 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

b) Central Mean Theorem

c) Centroid Limit Theorem

d) All of the mentioned

Answer: a) Central Limit Theorem

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

a) Modeling event/time data

b) Modeling bounded count data

c) Modeling contingency tables

d) All of the mentioned

Answer: b) Modeling bounded count data

4. Point out the correct statement.

a) The exponent of a normally distributed random variables follows what is called the log- normal distribution

b) Sums of normally distributed random variables are again normally distributed even if the variables are dependent

c) The square of a standard normal random variable follows what is called chi-squared distribution

d) All of the mentioned

Answer: d) All of the mentioned

5. _____ random variables are used to model rates.

a) Empirical

b) Binomial

c) Poisson

d) All of the mentioned

Answer: c) Poisson

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

a) True

b) False

Answer: b) False

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

a) Probability

b) Hypothesis

c) Causal

d) None of the mentioned

Answer: b) Hypothesis

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

a) 0

b) 5

c) 1

d) 10

Answer: a) 0

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

a) Outliers can have varying degrees of influence

b) Outliers can be the result of spurious or real processes

c) Outliers cannot conform to the regression relationship

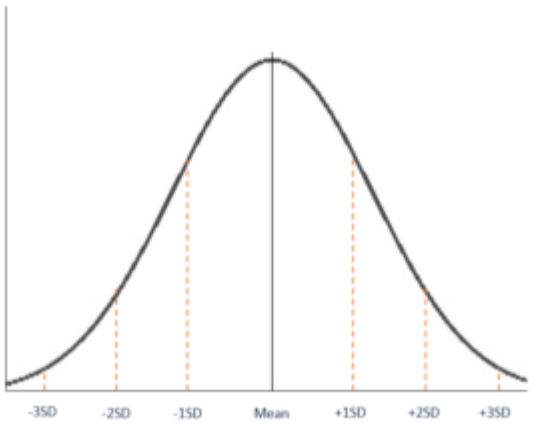
d) None of the mentioned

Answer: c) Outliers cannot conform to the regression relationship

Q10 and Q15 are subjective answer type questions, Answer them in your own words briefly.

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

Answer: The probability function that describes how the values of a variable are distributed. If in data distribution mean, median and mode are equal (i.e., mean = median = mode), we can say that the data is normally distributed. After plotting a bell curve of data, the area under curve to the left of mean and area under the curve to the right-hand side are equal, then the data is normally distributed.



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

Answer: There are many methods to deal with the missing values in dataset. Few of them are:

a) Simple Imputer
b) KNN Imputer
c) Iterative Imputer

Depending upon dataset requirement, we can select any imputer technique. Like if the data is continuous in nature, we can use simple imputer-mean function to fill missing values.

12. What is A/B testing?

Answer: A/B testing is a basic random control experiment. In this, we divide our dataset into two sets A & B. In A we don't do any change and in B we make suitable changes then we compare with users which model performs better.

It is a hypothetical testing methodology for making decisions that estimate population parameters based on sample statistics. The population refers to all the customers buying your product, while the sample refers to the number of customers that participated in the test.

13. Is mean imputation of missing data acceptable practice?

Answer: Yes, it can be accepted if missing values are lesser and it is a continuous data.

14. What is linear regression in statistics?

Answer: In this, we establish a relationship between two variables, one being scaler in nature and other can be dependent or independent variable. If we perform analysis with one variable it is called simple linear regression and if we perform analysis with more than one, it is called as multiple linear regression. This term is distinct from multivariate linear regression, where multiple correlated dependent variables are predicted, rather than a single scalar variable.

15. What are the various branches of statistics

Answer: There are two branches of statistics:

a) Descriptive Analysis – Deals with presentation and collection of data
b) Inferential Analysis -- It involves in drawing right conclusion from data.