ANSWERS FOR 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.

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

5. _____ random variables are used to model rates.

c) Poisson

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

b) False

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

b) Hypothesis

8. 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 or Gaussian distribution is a statistical tool represent the data in the form of Bell shaped curve. As the area under the curve is more at the centre, maximum data lies over there; i.e close to mean value of data set. This data has higher probability to occur. Opposite to this, data which further away from mean or central area has less probability. There is rule, which states that 68% of data falls in the region of first standard deviation, 95% in second standard deviation and 99.7% within third standard deviation, if the data is following perfect normal distribution.

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

Real time data sets may have missing data that can be either ignored or can be imputed using many techniques. Ignoring the data points with null values will results in possibility of losing important data. Thus imputing missing data with existing data the help of existing data is preferred. Techniques such as Mean/median imputation, KNN, Most frequent imputation, etc can be used.

12. What is A/B testing?

A/B testing is a tool used to compare two different forms of variable based on the user preferences.

13. Is mean imputation of missing data acceptable practice?

No, it is not preferable. Mean and Mode both are the central tendency of the data set, although it does not consider the correlation between two different variables. Same drawback observed in the mean imputation technique used for data imputation. Using this technique will narrow the variance because of smaller number of missing data set, which will ultimately reduce confidence interval.

14. What is linear regression in statistics?

Linear regression is a type of supervised machine learning algorithm used to determine the relationship between the input variables which is also called as features and output variables known as labels. It is used for prediction of continuous variables such as weather forecasting, market trend etc. Linear regression develops best fit line which minimizes the disparency between predicted and actual data set.

15. What are the various branches of statistics?

It is a science of collecting, organizing , and analyzing the data. Which is furthermore divided into two main categories Descriptive statistic and inferential statistics? Descriptive statistic is limited to organizing & summarizing data. Although inferential statistics is a tool where we use the available data that we have measured to form the conclusion.