STATISTICS WORKSHEET – 1

1. Bernoulli random variables take (only) the values 1 and 0.
   Ans. 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?
   Ans. a) Central Limit Theorem

3. Which of the following is incorrect with respect to use of Poisson distribution?
   Ans. b) Modelling 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 is also known as the Gaussian distribution. It is a distribution that is symmetric about the mean and shows that data near the mean are more frequent in occurrence than data far from the mean. In graph form, normal distribution will appear as a bell curve.
    i. In normal distribution the mean, mode and median are all equal.
    ii. In a normal distribution the mean is zero and the standard deviation is 1.

11. How do you handle missing data? What imputation techniques do you recommend?
    Ans. Missing data can be handled using various methods, which are:-
    i. Deleting the columns with missing data
    ii. Deleting the rows with missing data
    iii. Imputation - Filling the missing data with a value

iv. Imputation with an additional column – Adding column to identify whether the value has come from the original data or the imputed value.
v. Filling with a Regression Model

- **Imputation Techniques:-** Missing values in the data can be filled using various techniques, which are-
  i. Filling the missing data with the mean or median value if it's a numerical variable.
  ii. Filling the missing data with mode if it's a categorical value.
  iii. Filling the numerical value with 0 or -999, or some other number that will not occur in the data. This can be done so that the machine can recognize that the data is not real or is different.
  iv. Filling the categorical value with a new type for the missing values.

12. What is A/B testing?

Ans. A/B testing (also known as split testing or bucket testing) is a method of comparing two versions of a webpage or app against each other to determine which one performs better. A/B testing is essentially an experiment where two or more variants of a page are shown to users at random, and statistical analysis is used to determine which variation performs better for a given conversion goal. It helps you to determine exactly what factors or features effect the user experience in +ve way. Continuous A/B Testing can be used to improve the web page or app day to day. It lets you know whether your hypothesis is true or not i.e. whether there is a +ve result from the variations or not. If the first hypothesis fails you can continue iterating with different new hypothesis until you get the desired result.

13. Is mean imputation of missing data acceptable practice?

Ans. Mean Imputation of missing data can be done very easily but it is not an acceptable practice because:-
i. Mean Imputation ignores feature correlation b/w different columns.
ii. Mean Imputation reduces the variance of the data.

14. What is linear regression in statistics?

Ans. Linear regression is a kind of statistical analysis that attempts to show a relationship between two variables. Linear regression looks at various data points and plots a trend line. Linear regression is used to predict the value of a variable based on the value of another variable. The variable you want to predict is called the dependent variable. The variable you are using to predict the other variable's value is called the independent variable. Linear regression predicts the value of the dependent variable from the variable which shows the most linear relation with the dependent variable.

15. What are the various branches of statistics?

Ans. There are 2 branches of statistics, which are:-
i. Descriptive Statistics:
   a. Central Tendency – Mean, Median & Mode.
   b. Dispersion of Data – Range, Percentile, IQR, Standard Deviation, Skewness.

ii. Inferential Statistics: Used for analysis of data with large Population. A sample is taken out from that large population for its analysis & then that result in inferred to the larger population.