STATISTICS WORKSHEET-1

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

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.

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.

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?

Ans: The normal distribution curve is also known as bell shaped curve, because it resembles like a bell. The center of the bell shaped curve is the average of the statistical data. The width of the curve is given by the standard deviation. The broader the width of the curve the taller the curve is and vice-versa. For the normal distribution mean=median=mode.

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

Ans: One can delete the record of missing value, if we have huge data set.

But if data is small:

1. One can create a separate model to fill the missing values, use test-train technique. But it would be time taking process.

2. One of the best techniques is, one can simply enter the mean or median or mode of the data.

12. What is A/B testing?

Ans: A/B testing is the statistical way to compare two versions say version A or Version B. It also helps to understand the statistical significance of the two versions. It is randomized experiment to conduct two sample hypothesis testing. It is one of the the effective way to understand your customer.

13. Is mean imputation of missing data acceptable practice?

Ans: Imputation technique generally depends on the data set available. If the dataset is huge in number and have few missing values, one can remove the record.

But if the dataset have nearby values and not much extreme values mean imputation will be the better technique, and it will definitely give the unbiased results. But if there is lot of missing values so it may give bias result.

If the dataset has lot of extreme values than median imputation is beneficial, also if the frequency of particular data is high than mode imputation work better**.**

14. What is linear regression in statistics?

Ans: Linear regression is one of the regression technique (between independent and dependent variable), provided the dependent variable is continuous in nature.
The relation between independent and dependent is linear in nature or as the independent variable increases the dependent variable also increases.
Simple linear regression: It is method for predicting a quantitative response using single feature.

The equation is of form:
Y = a + bX
X= is feature
b is coefficient of x
a is intercept
Multiple linear regressions: It includes multiple features and creates a model to see the relation between those features and the label.
Each x represent different feature and each feature has its own coefficient. The feature has maximum coefficient, that variable has more impact in the model.

15. What are the various branches of statistics?

Ans:The two main branches of statistics are:
1. Descriptive Statistics: It is easy to describe and less in number. It helps to organize data and focuses on the main characteristics of the data. It provides a summary of data numerically or graphically.
2. Inferential Statistics: It generalizes the larger dataset and applies probability theory to draw conclusion. It allows to infer population parameters based on sample statistics and to model relationships with in the model.
……..