STATISTICS WORKSHEET – 1

Q1 to Q9 have only one correct answer. Choose the correct option to answer your questions

Q1. Bernoulli random variables take(only) the values 1 and 0.

Ans. a) True

Q2. 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

Q3. Which of the following is incorrect with respect to use of Poisson distribution?

Ans. b) Modeling bounded count data

Q4. 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.

Ans. D) All of the mentioned.

Q5. ________random variables are used to model rates.

Ans. C) Poisson distribution

Q6. Usually replacing the standard error by its estimated value does change the CLT.

Ans. B) False

Q7. Which of the following testing is concerned with making decisions using data?

Ans. b) Hypothesis testing

Q8. Normalized data are centered at _____ and have units equal to standard deviations of the original data.

Ans. a) 0

Q9. Which of the following statement is incorrect with respect to outliers?

Ans. c) Outliers cannot conform to the regression relationship

Q10 to Q15 are subjective answer type questions. Answer them in your words briefly.

Q10. What do you understand by the term Normal Distribution?

Ans. In normal distribution, data is distributed normally, it follows bell shaped curve.

If the data is left skewed or right skewed then convert into normal distribution.

The normal distribution is explained by two parameters Mean & Standard deviation. The normal distribution with Mean=0 and Standard deviation=1 is called as Standard normal distribution.

Q11. How do you handle missing data? What imputation techniques do you recommend?

Ans. We use two methods for handling the missing data, imputation and removal of data.

Its depends only data which technique is good for our model.

Usually, missing data is found in form of 'NAN', there are different kind of imputers to solve this problem.

a) *<u>Simple imputer</u>* – It works on the basic mean method for imputing the values.

b) *<u>KNN imputer</u>* - Alike KNN algorithm this imputer finds the two closest neighbors but the important part to remember is you have to pass continuous data along with the column where you want to impute the values, here it will find the number of neighbors (as mentioned in code example neighbors=2) and it will take the average of two close neighbors to fill the value.

c)*<u>Iterative imputer</u>* – This imputer works on prediction basis here you need to pass the continuous data and the outcome will be predicted by this imputer itself, this imputer internally predicts the outcome to fill NAN values

Q12. What is A/B testing?

Ans. A/B testing is a basic randomized control experiment. It is a way to compare the two Versions of a variable to find out which one is better.

You own a company and want to increase the sales of product. Here, either you can Use random experiments, or you can apply scientific and statistical methods.

A/B testing is one of the most prominent and widely used statistical tool.

Q13. Is mean imputation of missing data acceptable practice?

Ans. Mean imputation is typically considered terrible practice since it ignores feature Correlation. Using mean imputation technique, it distorts the relationship between Variables. There is a possibility of biasness which affects the model confidence.

Q14. What is linear regression in statistics?

Ans. Linear regression provide the relationship between two continuous variables.

A linear regression model predicts the dependent variable using a regression line Based on the independent variable.

Q15. What are the various branches of statistics?

Ans. Two main branches of statistics:

a) *Descriptive Statistics* –

It defines the data where we are able to describe by using

Measure of Central Tendency – Mean, median, mode

Measure of Dispersion- Variance and Standard deviation, here you can find how much data is varying

b) *Inferential statistics* – This statistic is purely based on sampling here the essence of central limit theorem plays a vital role and says that if you take any sample from your normal distribution and you plot the same it will also form normal distribution. It also says that, mean of sample mean = Mean of population mean