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) Modelling 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.

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

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

Ans: The normal distribution is the most important probability distribution in statistics for independent, random variables. It is a bell-shaped curve where most of the observations cluster around the central peak, and the probabilities for values further away from the mean taper off equally in both directions. Extreme values in both tails of the distribution are similarly unlikely. The area under the whole curve is equal to 1, or 100% and at the centre let's say (u) Mean = Median = Mode.

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

Ans: There are various imputation techniques to tackle missing data such as Normal Imputation, Imputation based on class label, Model based imputation technique but in statistics we mainly use normal imputation technique. Here we basically replace the missing data with either mean, median or mode. If the data is numerical, we can use mean and median values to replace else if the data is categorical, we can use mode which is a frequently occurring value.

13.What is A/B testing?

Ans: A/B testing is a form of statistical and two-sample hypothesis testing. Statistical hypothesis testing is a method in which a sample dataset is compared against the population data. Two-sample hypothesis testing is a method in determining whether the differences between the two samples are statistically significant or not.

13. Is mean imputation of missing data acceptable practice?

Ans: Mean imputation is generally considered to be a bad practice because it doesn't take into account feature correlation. For example, imagine we have a table showing age and fitness score and imagine that an eighty-year-old has a missing fitness score. If we took the average fitness score from an age range of 15 to 80, then the eighty-year-old will appear to have a much higher fitness score that he actually should. Second, mean imputation reduces the variance of the data and increases bias in our data. This leads to a less accurate model and a narrower confidence interval due to a smaller variance.

14. What is linear regression in statistics?

Ans: Linear regression is a basic and commonly used type of predictive analysis. Regression estimates are used to explain the relationship between one dependent variable and one or more independent variables. The simplest form of the regression equation with one dependent and one independent variable is defined by the formula y = c + m*x, where y = estimated dependent variable score, c = constant, m= regression coefficient, and x = score on the independent variable. Three major uses for regression analysis are determining the strength of predictors, forecasting an effect, and trend forecasting.

15. What are the various branches of statistics?
Ans: There are mainly two branches of statistics Descriptive and Inferential.

Descriptive statistics is the first part of statistics that deals with the collection of data. It again has two parts measure of central tendency and measure of dispersion.

Inference statistics are techniques that enable statisticians to use the information collected from the sample to conclude, bring decisions, or predict a defined population. There are different types of inferential statistics but me mostly use t-test and ANOVA.