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) Modeling 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. 10. Usually replaacing the standard error by its estimated value does change the CLT.

Ans.- (b) False

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

Ans.- (b) Hypothesis

8. 4. 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. - A normal distribution is a type of continuous probability distribution in which most data points cluster toward the middle of the range, while the rest taper off symmetrically toward either extreme. The middle of the range is also known as the mean of the distribution.

The normal distribution is also known as a Gaussian distribution

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

Ans.- Here are the most common ways of handling missing data

Zero Replacement: Here, you replace the missing value with zero irrespective of everything.

Min or Max Replacement: Replace the missing value with the minimum or maximum value of a feature.

Mean/ Median/ Mode Replacement: Replace missing value with mean or median or most frequent feature value.

Also, one can replace the value of the missing cell with the previous cell's value. This kind of technique is popular while inputting time series data. For example, if the price of an instrument is missing on the i-th day, it makes sense to replace it with the (i-1)-th day's price.

12. What is A/B testing?

Ans.- A/B testing, at its most basic, is a way to compare two versions of something to figure out which performs better. While it's most often associated with websites and apps, Fung says the method is almost 100 years old.

In the 1920s statistician and biologist Ronald Fisher discovered the most important principles behind A/B testing and randomized controlled experiments in general. "He wasn't the first to run an experiment like this, but he was the first to figure out the basic principles and mathematics and make them a science," Fung says.

Fisher ran agricultural experiments, asking questions such as, What happens if I put more fertilizer on this land? The principles persisted and in the early 1950s scientists started running clinical trials in medicine. In the 1960s and 1970s the concept was adapted by marketers to evaluate direct response campaigns (e.g., would a postcard or a letter to target customers result in more sales?).

A/B testing, in its current form, came into existence in the 1990s. Fung says that throughout the past century the math behind the tests hasn't changed. "It's the same core concepts, but now you're doing it online, in a real-time environment, and on a different scale in terms of number of participants and number of experiments."

13. Is mean imputation of missing data acceptable practice?

Ans.- The process of replacing null values in a data collection with the data's mean is known as mean imputation.

Mean imputation is typically considered terrible practice since it ignores feature correlation. Consider the following scenario: we have a table with age and fitness scores, and an eight-year-old has a missing fitness score. If we average the fitness scores of people between the ages of 15 and 80, the eighty-year-old will appear to have a significantly greater fitness level than he actually does.

Second, mean imputation decreases the variance of our data while increasing bias. As a result of the reduced variance, the model is less accurate and the confidence interval is narrower.

14. What is linear regression in statistics?

Ans.- Linear regression analysis 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.

This form of analysis estimates the coefficients of the linear equation, involving one or more independent variables that best predict the value of the dependent variable. Linear regression fits a straight line or surface that minimizes the discrepancies between predicted and actual output values. There are simple linear regression calculators that use a "least squares" method

to discover the best-fit line for a set of paired data. You then estimate the value of X (dependent variable) from Y (independent variable)

15. What are the various branches of statistics?

Ans.- Statistics is a study of presentation, analysis, collection, interpretation and organization of data

There are two main branches of statistics

- Inferential Statistic.

- Descriptive Statistic.

Inferential Statistics:

Inferential statistics used to make inference and describe about the population. These stats are more useful when its not easy or possible to examine each member of the population.

Descriptive Statistics:

Descriptive statistics are use to get a brief summary of data. You can have the summary of data in numerical or graphycal form