STATISTICS WORKSHEET -1

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

Answer: - 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?

Answer: - a) Central Limit Theorem.

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

Answer: - b) Modeling bounded count data

4. Point out the correct statement.

Answer: - c) The square of a standard normal random variable follows what is called chi-squared distribution.

5. _____ random variables are used to model rates.

Answer: - c) Poisson

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

Answer: - False

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

Answer: - b) Hypothesis

8. Normalized data are centered at_____and have units equal to standard deviations of the original data.

Answer: - a) 0

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

Answer: - c) Outliers cannot conform to the regression relationship.

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

Answer: - Normal distribution is a type of continuous probability distribution that is symmetrical and bell-shaped, where most of the data points cluster around the mean,

with fewer points appearing as you move away from the mean in both directions. The mean, median, and mode of a normal distribution are all equal, and the distribution is defined by its mean (center) and standard deviation (spread). It is often used in statistics because many natural phenomena, like heights or test scores, tend to follow this pattern, making it a key concept in probability and inferential statistics.

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

Answer: - There are few imputation techniques which we can do. **Mean/median/Mode** Imputation. With the help of Mean/Median/mode, we can replace the missing values with the mean for numerical data, median (useful for skewed distributions) or Mode (For categorical data).

There is also KNN algorithm with the help of KNN we can impute missing values based on similarity of other data points. And one of the algo Regression Imputation. Use regression models to predict and fill the missing values based on available data.

There are few more algo ways to impute the missing vales from data.

We can also Remove the missing Values from data, this technique is simple it can lead to loss the data. This is not the right choice to delete data to handle missing values from data. However, its not good to delete the data which missing in columns. I never prefer to delete it.

12. What is A/B testing?

Answer: - A/B testing is a powerful tool that enables organizations to make evidence-based decisions by comparing two versions of a variable and measuring their impact on specific outcomes. It's widely used across various industries to enhance user experience, increase conversion rates, and ultimately drive business growth.

A/B testing is also known as split testing. This is statistical method used to compare two version of a variable to determine which one perform better in achieving a specific outcome.

Benefits of A/B testing. It allows businesses to make informed decisions based on user behavior rather than assumptions.

Minimized Risk. Testing changes on a smaller segment of users before full implementation reduces the risk of negatively impacting overall performance.

Optimization. Continues A/B testing helps in optimizing products, website, and marketing strategies to improve user experience and increase conversions.

13. Is mean imputation of missing data acceptable practice?

Answer: - Mean imputation is simple and commonly used technique but it has many drawbacks and some time it is not good to use because Mean imputation reduces the Variability in the dataset. This can also distort the true variance and effect statistical inferences.

Mean imputation can lead to Bias estimate of relationship between variables, particularly in regression models. Because it does not consider the underlying patterns or relationship that might exist in the missing data.

Ignoring the pattern of missingness. If the data are missing in way that is not random data missing due to some underlying reason. Mean imputation ignores this pattern and assumes data are missing completely at random, which can lead misleading conclusions. As per me this is not good to use mean data. **KNN** is imputation good to use. As they tend to preserve the structure and relationships within the data better.

14. What is linear regression in statistics?

Answer: - Linear regression is statistical method used to model and analyze the relationship between two variables by fitting a linear equation to the observed data. The goal is to predict the value of a dependent variable (often denoted as y) based on one or more independent variables (often denoted as x).

It is commonly used for prediction and to understand how changes in the variables affect the outcome. Multiple linear regression extends this concept to more than one predictor variable.

15. What are the various branches of statistics?

Answer: - In statistics there is branches 2 branches **Descriptive Statistics and Inferential Statistics,** both are focuses on different aspects of analyzing and interpreting data.

Descriptive Statistics is to summarize, organize, and describe data in a meaningful way.

Inferential Statistics is to make inferences, predictions, or generalizations about a population based on a sample.

These branches are backbone of statistical analysis, helping researchers and data scientists make sense of complex data sets and derive insights for decision-making.