STATISTICS WORKSHEET

C SOMA SEKHAR GOUD

BATCH: DS2307

1. a) True

2. a) Central Limit Theorem

3. b) Modeling bounded count data

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

5. c) Poisson

6. b) False

7. b) Hypothesis

8. a) 0

9. c) Outliers cannot conform to the regression relationship

10.

Normal distribution, also known as Gaussian distribution, is a fundamental concept in statistics and probability. It describes a symmetric bell-shaped curve that is characterized by its mean (average) and standard deviation. In a normal distribution, the majority of data points cluster around the mean, and the further away from the mean a value is, the less likely it is to occur. Many natural phenomena, such as heights, weights, and errors in measurements, tend to follow a normal distribution. This distribution is crucial in statistical analysis because it allows us to make predictions and infer probabilities about data.

11. Dealing with missing data is a common challenge in data analysis. There are few techniques that can be used to estimate missing values based on available information. Mean imputation (replacing missing values with the mean of the non-missing values), Median imputation (using the median of non-missing values),

And predictive modeling methods like regression imputation and k-nearest neighbours (KNN) imputation.

The choice of technique depends on the nature of the data and the analysis goals.

12.

A/B testing is a technique used to evaluate and improve machine learning models which involves in experimentation to compare two versions of something to determine which one performs better. In this methodology, a sample population is partitioned into two groups, wherein one group is exposed to the original version (A) while the other encounters a modified version (B). By tracking and analyzing user interactions or outcomes, A/B testing gauges the impact of the modification. This process allows machine learning practitioners to objectively evaluate which variant yields superior results, thus informing data-driven decisions about model enhancements or feature changes. Through its iterative and data-centric nature, A/B testing empowers the fine-tuning and optimization of machine learning models, enabling their alignment with real-world requirements and user preferences.

13.

The process of replacing null values in a data collection with the data's mean is known as mean imputation. While mean imputation is a simple and quick method, it has limitations. Mean imputation ignores feature correlation. It 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.

Linear regression is a statistical technique used to model the relationship between a dependent variable and one or more independent variables by fitting a linear equation to the observed data. The goal is to find the best-fitting line (linear equation) that minimizes the differences between the observed values and the predicted values generated by the model. It is widely used for prediction and understanding the influence of one or more independent variables on the dependent variable.

15.

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

There are two main branches of statistics

- Inferential Statistic. (Making predictions and drawing conclusions about populations based on samples.)

- Descriptive Statistic. (Summarizing and describing data using measures like mean, median, and standard deviation.)