STATISTICS WORKSHEET-1

1. a) True
2. a) Central Limited Theorem
3. b) Modeling bounded count data
4. b) Sums of normally distributed random variables are again normally distributed even if the variables are dependent
5. c) Poisson
6. b) False
7. b) Hypothesis
8. a) 0
9. c) Outliers cannot conform to the regression relationship
10. Normal Distribution is a bell-shaped frequency distribution curve which helps describe all the possible values a random variable can take within a given range with most of the distribution area is in the middle and few are in the tails, at the extremes.

11. Missing data can be handled in a variety of ways, depending on the context and the type of data. Generally, the most common approach is to use imputation techniques to fill in the missing values.

    The most common imputation techniques are mean/mode imputation, k-nearest neighbor imputation, and multiple imputation.

    Mean/mode imputation involves replacing missing values with the mean or mode of the available data. This is a simple and straightforward approach, but it can lead to biased results if the data is not normally distributed.

    K-nearest neighbor imputation is a more sophisticated approach that uses the values of the nearest neighbors to fill in the missing values. This

technique is more accurate than mean/mode imputation, but it can be computationally expensive.

Multiple imputation is a technique that involves creating multiple datasets with different imputed values for the missing data. This approach is more accurate than the other two, but it is also more computationally expensive.

No matter which imputation technique is used, it is important to remember that the results should be interpreted with caution, as the imputed values may not accurately reflect the true values.

12. A/B testing is a statistical method used in data science to compare two versions of a product or service to determine which one performs better. It involves randomly dividing a sample population into two groups, where one group is exposed to the original version (control group) and the other group is exposed to a modified version (treatment group).

The performance of each group is then measured and compared to determine which version is more effective. A/B testing is commonly used in marketing, web design, and product development to optimize user experience and increase conversion rates.

13. Mean imputation is a common method for handling missing data in machine learning. It involves replacing missing values with the mean value of the feature. While it is a simple and easy-to-implement method, it has some limitations and potential drawbacks.

One of the main issues with mean imputation is that it can lead to biased estimates of the variance and covariance of the data. This is because the imputed values are all the same, which can artificially reduce the variability of the data. Additionally, mean imputation assumes that the missing values are missing completely at random (MCAR), which may not always be the case.

It is important to carefully consider the implications of using mean imputation and to explore alternative methods when appropriate.

14. Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables. It is a technique used to predict the value of a dependent variable based on the values of one or more independent variables. The goal of linear regression is to find the best-fit line that represents the relationship between the variables. The line is determined by minimizing the sum of the squared differences between the predicted values and the actual values.

15. Statistic is the study of presentation, analysis, collection, interpretation and organization of data.

There are two branches of statistics:

- Descriptive Statistics.
- Inferential Statistics.

Descriptive Statistics:

In this type of statistics, the data is summarized through the given observation. The summarization is one from a sample of population using parameters such as the mean or standard deviation.

Descriptive statistics are also categorized into four different categories:

- Measure of central tendency
- Measures of variance.

Inferential Statistics:

Inferential statistics is a branch of statistics that involves using sample data to make inferences or draw conclusions about a larger population. It is used to test hypotheses, estimate population parameters, and make predictions based on the data collected from a sample.

The different types of calculation of inferential statistics include:

- Regression analysis
- Analysis of variance (ANOVA)
- Analysis of covariance (ANCOVA)
- Statistical significance (t-test)
- Correlation analysis