STATISTICS WORKSHEET-1

1. A) True
2. A) The Central Limit Theorem.
3. B) Modeling bounded count data.
4. D) All of the mentioned
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 type of probability distribution that is symmetrical around the mean. It is also known as a Gaussian distribution and is commonly used to describe real-world phenomena such as height, weight, or test scores. A normal distribution is characterized by the bell-shaped curve, which shows that the probability of an event occurring decreases as it moves away from the mean.
11. Missing data can be handled in a variety of ways depending on the context. One of the most common methods of handling missing data is to use imputation techniques. Imputation is the process of replacing missing values with estimated ones. Common imputation techniques include mean imputation, median imputation, and mode imputation. Mean imputation replaces missing values with the mean of the rest of the data. Median imputation replaces missing values with the median of the rest of the data. Mode imputation replaces missing values with the mode of the rest of the data. Other imputation techniques include k-nearest neighbors imputation, multiple imputations, and predictive modeling. K-nearest neighbors imputation uses the values of the nearest neighbors to a missing value to estimate the missing data. Multiple imputations create multiple datasets with different imputations and then analyze the results to determine which imputation was the most accurate. Predictive modeling uses machine learning algorithms to predict the missing values. No matter which imputation technique is used, it is important to consider the context and which technique is most appropriate. It is also important to consider the impact of imputation on the results. Imputation can introduce bias and distortion, so it is important to use**.**

12. A/B testing is a type of experiment in which you split your web traffic or user base into two groups, and show two different versions of a web page, app, email, and so on, with the goal of comparing the results to find the more successful version.

13. 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. 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. Statistics is the branch of mathematics that deals with data. Data (technically a plural word; the singular is 'datum') is a collection of values. For most of what we do, it will be numerical data (such as the inflation rate, the number of bees in a colony, or the marks in a class test), but it can also take other forms (such as the political party a voter intends to vote for, the football team they support, and so on).

    The Branches of Statistics

    Two branches, *descriptive statistics* and *inferential statistics*, comprise the field of statistics.

Descriptive statistics:

This is a branch of statistics which deals with methods of collection of data, its presentation and organization in various forms, such as distribution tables, graphs (e.g., ogive, Lorenz curves, etc.), diagrams (e.g., pie charts) and finding measures of central tendency and measures of dispersion or spread which are used in the description of data. Managers, CEOs. etc. make use of descriptive statistics in presenting their annual reports, financial accounts and bank statements.

Descriptive statistics is used to present the data in an understandable way, so that a meaningful description can be made.

Inferential or predictive statistics:

This is a branch of statistics which deals with techniques used for analysis of data, making estimates that lead to predictions and drawing conclusions or inferences from

limited information taken on sample basis and testing the reliability of the estimates or predictions.
Inferential statistics is used to make comparisons or predictions about a larger group, known as population, using information gathered about a small part of that population called a sample.

Inferential statistics answers questions, such as "what is this data telling us about?" and "what should we do?" Techniques used are forecasting trends, hypothesis testing, kurtosis, skewness, etc.