Statistics Worksheet -1

1. A. True
2. A. Central Limit Theorem
3. B. Modeling bounded count data
4. D. All the above
5. C. Poisson
6. B. False
7. B. Hypothesis
8. A. 0
9. C. Outliers cannot conform toy he regression relationship
10. A normal distribution, also known as a Gaussian distribution, is a probability distribution that is symmetric about the mean, indicating that data near the mean are more frequent in occurrence than data far from the mean. In graphical form, the normal distribution will appear as a bell curve. It is characterized by its mean (average) and standard deviation (spread or width).
11. 1. **Deletion Methods:** Removing rows with missing values (listwise deletion) or only the specific cells with missing data (pairwise deletion).

2. **Imputation Methods:** Replacing missing values with substituted values, such as:

- **Mean/Median Imputation:** Replacing missing values with the mean or median of the column.
- **Mode Imputation:** Using the most frequent value in the column.
- **Regression Imputation:** Predicting the missing values using a regression model.
- **Multiple Imputation:** Using multiple datasets to account for the uncertainty around missing data.
- **K-Nearest Neighbors (KNN):** Using the K-nearest neighbors to predict and impute the missing values.

12. A/B testing, or split testing, is a statistical method used to compare two versions of a webpage, product, or feature to determine which one performs better. By randomly splitting users into two groups (A and B) and exposing each group to a different version, analysts can measure and compare the effects of the changes made to key performance indicators (KPIs) like click-through rates, conversion rates, and user engagement.
13. Yes, but not for all types of datasets. Mean imputation is a simple method for handling missing data but can lead to biased estimates and reduced variability in the data. While it may be acceptable for small amounts of missing data, it is generally not recommended for large datasets or when the missing data mechanism is not random. More sophisticated methods like multiple imputation or model-based imputation are often preferred.
14. Linear regression is a statistical technique used to model and analyze the relationships between a dependent variable and one or more independent variables. The goal is to find the linear equation that best predicts the dependent variable based on the independent variables.

The equation $y=\beta0+\beta1x1+\beta2x2+\ldots+\beta nxn+\epsilon$, where β are the coefficients and ϵ is the error.

15. **Descriptive Statistics:** Summarizing and describing the features of a dataset.
    Inferential Statistics: Making inferences and predictions about a population based on a sample of data.
    Probability Theory: Studying randomness and uncertainty in various processes.
    Multivariate Statistics: Analyzing data that involves multiple variables to understand relationships and patterns.
    Exploratory Data Analysis (EDA): Analyzing data sets to summarize their main characteristics often with visual methods.