STATISTICS WORKSHEET-1

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

a) True

b) False

Ans:- b) False

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?

a) Central Limit Theorem

b) Central Mean Theorem

c) Centroid Limit Theorem

d) All of the mentioned

Ans:- a) Central Limit Theorem

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

a) Modeling event/time data

b) Modeling bounded count data

c) Modeling contingency tables

d) All of the mentioned

Ans:- b) Modeling bounded count data

4. Point out the correct statement.

a) The exponent of a normally distributed random variables follows what is called the log- normal distribution

b) Sums of normally distributed random variables are again normally distributed even if the variables are dependent

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

d) All of the mentioned

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

5. ______ random variables are used to model rates.

a) Empirical

b) Binomial

c) Poisson

d) All of the mentioned

Ans:- c) Poisson

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

a) True

b) False

Ans:- b) False

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

a) Probability

b) Hypothesis

c) Causal

d) None of the mentioned

Ans:- b) Hypothesis

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

a) 0

b) 5

c) 1

d) 10

Ans:- a) 0

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

a) Outliers can have varying degrees of influence

b) Outliers can be the result of spurious or real processes

c) Outliers cannot conform to the regression relationship

d) None of the mentioned

Ans:- c) Outliers cannot conform to the regression relationship

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

Ans:- The normal distribution, also known as the Gaussian distribution, is a fundamental concept in statistics and probability theory. It describes a symmetrical, bell-shaped curve that is characterized by two parameters: the mean and the standard deviation. In a normal distribution the mean represents the center of the distribution and also corresponds to the highest point on the curve. The standard deviation determines the spread or dispersion of the data points around the mean. A larger standard deviation indicates greater variability in the data, resulting in a wider and flatter curve, while a smaller standard deviation results in a narrower and taller curve.

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

Ans:- Handling missing data is a crucial aspect of data analysis, and there are several techniques for imputing missing values. The choice of imputation method depends on the nature of the data and the specific characteristics of the dataset.

Here are some common techniques:

Mean/Median/Mode Imputation: This involves replacing missing values with the mean, median, or mode of the observed data for that variable. It's a simple and quick method but may not capture the true variability of the data.

Forward Fill/Backward Fill: In time series data, missing values can be filled with the most recent non-missing value (forward fill) or the next non-missing value (backward fill). This method is useful when missing values occur in sequences.

Linear Interpolation: Missing values can be estimated by linearly interpolating between adjacent observed values. This method assumes a linear relationship between the observed values.

Multiple Imputation: This technique involves creating multiple imputed datasets by estimating missing values based on observed data and their uncertainty. Statistical methods such as regression models or predictive mean matching can be used for imputation.

K-Nearest Neighbors (KNN) Imputation: Missing values are imputed based on the values of nearest neighbors in the feature space. This method considers the similarity between observations to estimate missing values.

Expectation-Maximization (EM) Algorithm: EM algorithm iteratively estimates missing values by maximizing the likelihood of observed data. It's particularly useful for datasets with complex patterns of missingness.

Matrix Factorization: This method decomposes the dataset into lower-dimensional matrices and estimates missing values based on the relationships between observed values.

12. What is A/B testing?

Ans:- A/B testing, also known as split testing, is a statistical method used to compare two or more versions of a webpage, app, email, or other marketing asset to determine which one performs better. In an A/B test, users are randomly assigned to different variations, with each group experiencing a different version of the asset. By analyzing the performance metrics, such as click-through rates, conversion rates, or revenue, between the groups, one can determine which version is more effective in achieving the desired outcome. A/B testing helps businesses make data-driven decisions by providing insights into user preferences, behaviors, and preferences, ultimately leading to improved performance and optimization of marketing strategies and user experiences.

13. Is mean imputation of missing data acceptable practice?

Ans:- Mean imputation is a common technique for handling missing data because of its simplicity. It involves replacing missing values with the mean of the observed data for that variable. While it's easy to implement and maintains the overall mean of the dataset, there are some concerns. For instance, mean imputation doesn't consider the uncertainty introduced by imputing missing values, which can lead to underestimating the variability of the data. Additionally, it assumes that missing values are missing completely at random (MCAR), which might not always be the case in real-world scenarios. Despite these limitations, mean imputation can still be suitable in certain situations, especially when the amount of missing data is small and missingness is relatively random. However, it's important to acknowledge its limitations and consider alternative imputation methods when necessary, such as multiple imputation or model-based imputation, to address more complex missing data patterns and assumptions.

14. What is linear regression in statistics?

Ans:- In statistics, linear regression is a method used to analyze the relationship between two or more variables. It's like drawing a straight line through a cloud of points on a graph. This line represents the best-fit relationship between the variables. One variable is considered the dependent variable, while the others are independent variables. For example, in studying how study time affects exam scores, study time would be the independent variable, and exam scores would be the dependent variable. The goal is to find the equation of the line that best fits the data, allowing us to predict one variable based on the other. Linear regression assumes that the relationship between the variables is linear, meaning that as one variable changes, the other changes at a constant rate. It's a widely used technique in fields such as economics, sociology, and psychology for understanding and predicting outcomes based on observed data.

15. What are the various branches of statistics

Ans:- tatistics is a broad field that encompasses various branches, each focusing on different aspects of data analysis, interpretation, and application.

Some of the main branches of statistics are:

Descriptive Statistics: Summarizes and organizes data, including measures like averages and charts.

Inferential Statistics: Draws conclusions or predictions about populations based on sample data.

Probability Theory: Studies the likelihood of events occurring, forming the basis of statistical methods.

Bayesian Statistics: Updates beliefs about parameters or hypotheses using prior knowledge and observed data.

Nonparametric Statistics: Analyzes data without specific distributional assumptions, offering versatility.

Biostatistics: Applies statistical methods to biological and medical data, aiding research and decision-making in healthcare.

Econometrics: Analyzes economic data to uncover relationships and make forecasts.

Multivariate Statistics: Examines datasets with multiple variables to reveal complex relationships.

Spatial Statistics: Focuses on data with geographical components, exploring spatial patterns and relationships.