STATISTICS WORKSHEET-1

Q1 to Q9 have only one correct answer. Choose the correct option to answer your question.

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

a) True

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

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

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

5. _____ random variables are used to model rates.

a) Empirical

b) Binomial

c) Poisson

d) All of the mentioned

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

a) True

b) False

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

a) Probability

b) Hypothesis

c) Causal

d) None of the mentioned

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

a) 0

b) 5

c) 1

d) 10

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

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

ANS- A key idea in statistics, the normal distribution describes how data points are dispersed in a bell-shaped, symmetrical curve. The salient features and attributes of a normal distribution are as follows:

Characteristics of Normal Distribution:

1. **Symmetry**: The distribution is symmetric about its mean. This means that the left and right sides of the curve are mirror images of each other.
2. **Mean, Median, and Mode**: In a normal distribution, the mean (average), median (middle value), and mode (most frequent value) are all equal and located at the centre of the distribution.
3. **Bell-shaped Curve**: The graph of a normal distribution is bell-shaped, with the highest point at the mean and tails that approach, but never touch, the horizontal axis.
4. **Empirical Rule**: Approximately:
    - 68% of the data falls within one standard deviation of the mean.
    - 95% of the data falls within two standard deviations.
    - 99.7% of the data falls within three standard deviations.

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

ANS- Handling missing data is a crucial step in the data preprocessing phase of any data analysis or machine learning project. The approach you choose depends on the nature of the data, the amount of missingness, and the specific requirements of your analysis. Here are some common strategies and imputation techniques for handling missing data:

Handling Missing Data

1. **Remove Missing Data**:
2. **Imputation Techniques**:
    - **Mean/Median/Mode Imputation**:
        - **Mean**: Replace missing values with the mean of the column (useful for numerical data).
        - **Median**: Replace missing values with the median (more robust to outliers).
        - **Mode**: Replace missing values with the mode (for categorical data).
    - **Forward/Backward Fill**:
        - **Forward Fill**: Replace missing values with the last valid observation (useful in time series).
        - **Backward Fill**: Replace missing values with the next valid observation.

- **K-Nearest Neighbors (KNN) Imputation**
- **Regression Imputation**
- **Multiple Imputation**
- **Interpolation**
- **Using Algorithms that Handle Missing Data**

Handling missing data effectively requires careful consideration of the data and the analysis goals. Imputation techniques can help maintain the integrity of the dataset, but they should be chosen based on the specific characteristics of the data and the analysis context.

12. What is A/B testing?

ANS- A statistical technique called A/B testing, or split testing, compares two versions of a single variable to see which one works better. Making data-driven judgements is a common practice in marketing, web design, product development, and other domains.

A/B testing is a powerful tool for optimizing various aspects of business operations, marketing strategies, and product development. By comparing the effectiveness of two versions, organizations can make informed decisions that enhance performance and drive better results.

13. Is mean imputation of missing data acceptable practice?

ANS- A popular method for dealing with missing data is mean imputation, in which the mean of the observed values for a variable is used to fill in the missing values in a dataset. Although it is simple and quick to use, there are a number of significant drawbacks, and it might not always be the optimal course of action.

While mean imputation can be a quick and easy solution for handling missing data, it is generally not recommended due to its potential to introduce bias and distort variability. More sophisticated methods that account for the underlying data distribution and relationships between variables are often preferred for better accuracy and reliability in statistical analyses. The choice of method should be guided by the nature of the data, the amount and mechanism of the missingness, and the specific goals of the analysis.

14. What is linear regression in statistics?

ANS- A basic statistical technique called linear regression is used to describe the relationship between one or more independent variables, sometimes referred to as predictor variables, and a dependent variable, also called the response variable. Finding the best-fitting straight line to represent how the dependent variable varies as a function of the independent factors is the aim of linear regression, or the hyperplane in multiple regression.

Linear regression is a powerful and widely used statistical technique that provides a straightforward way to model relationships between variables. Its simplicity and interpretability make it an essential tool in data analysis, economics, social sciences, and many other fields. However, it is crucial to ensure that the underlying assumptions are met to obtain valid and reliable results.

15. What are the various branches of statistics

ANS- Statistics is a broad field that encompasses various branches, each focusing on different aspects of data collection, analysis, interpretation, and presentation. Here are the primary branches of statistics:

1. Descriptive Statistics- This branch deals with summarizing and organizing data. It provides simple summaries about the sample and measures.

2. Inferential Statistics- Inferential statistics involves using a random sample of data to make inferences about the larger population from which the sample is drawn.

3. Parametric Statistics- This branch assumes that the data follows a certain distribution (usually normal distribution) and relies on parameters like mean and standard deviation.

4. Non-Parametric Statistics- Non-parametric statistics do not assume a specific distribution for the data. They are useful for analyzing data that do not meet the assumptions required for parametric tests.

5. Biostatistics-This branch applies statistical methods to the field of biology, particularly in the design and analysis of experiments in health sciences and medicine.

6. Econometrics

7. Quality Control and Industrial Statistics