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) Modelling event/time data

b) Modelling bounded count data

c) Modelling 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. Usually replacing the standard error by its estimated value does change the CLT.

a) True

b) False

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

a) Probability

b) Hypothesis

c) Causal

d) None of the mentioned

7. Normalized data are centred at_____and have units equal to standard deviations of the original data.

a) 0

b) 5

c) 1

d) 10

8. 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

WORKSHEET

Q10and Q15 are subjective answer type questions, Answer them in your own words briefly.

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

Answer: The Normal Distribution, also known as the Gaussian distribution, is a fundamental concept in statistics and probability theory. It describes how the values of a random variable are distributed and is characterized by its bell-shaped curve.

The Normal Distribution is a continuous probability distribution defined by its mean (μ) and standard deviation (σ). It is symmetric about its mean, meaning that the left and right sides of the distribution are mirror images of each other.

The Normal Distribution is widely used in various fields including:

- **Statistics:** It forms the basis for many statistical tests and methods.
- **Finance:** To model stock returns and assess risk.
- **Natural Sciences:** To describe natural phenomena like heights, weights, and measurement errors.
- **Quality Control:** To monitor and control process variations.

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

Answer: Handling missing data is a crucial step in data preprocessing, as it can significantly affect the quality of the analysis and the performance of machine learning models. There are several techniques for dealing with missing data, each with its own advantages and limitations. Here's a comprehensive guide to handling missing data and some recommended imputation techniques:

1. Identify Missing Data

- **Types of Missing Data:**
    - **Missing Completely at Random (MCAR):** Missingness is independent of both observed and unobserved data.
    - **Missing at Random (MAR):** Missingness is related to observed data but not to the missing values themselves.
    - **Missing Not at Random (MNAR):** Missingness is related to the missing values themselves.

2. Strategies for Handling Missing Data

a) Remove Missing Data

- **Listwise Deletion:** Remove any observation with one or more missing values. Suitable when missing data is small in proportion and random.

- **Pairwise Deletion:** Use available data for each analysis, including only those cases where the necessary variables are present. Useful for correlation and covariance calculations.

b) Imputation Techniques

1. **Simple Imputation**

- **Mean Imputation:** Replace missing values with the mean of the observed values for that variable. Suitable for numerical data but can reduce variability.
- **Median Imputation:** Replace missing values with the median. Robust to outliers and useful for skewed distributions.
- **Mode Imputation:** Replace missing values with the most frequent value. Commonly used for categorical data.

2. **Advanced Imputation Techniques**

- **K-Nearest Neighbours (KNN) Imputation:** Impute missing values using the values from the nearest neighbours based on a distance metric. Effective but computationally intensive.
- **Regression Imputation:** Predict missing values using a regression model based on other variables. Suitable for numerical data and assumes linear relationships.
- **Multiple Imputation:** Generate multiple datasets with different imputed values and combine results. This approach accounts for the uncertainty in missing data imputation and provides more robust estimates.
- **Expectation-Maximization (EM):** An iterative approach that estimates missing data using maximum likelihood estimation. Suitable for complex datasets and models.

3. **Model-Based Imputation**

- **Using Machine Learning Models:** Train models (e.g., Random Forests, Gradient Boosting) to predict missing values based on other variables. Useful for capturing complex relationships.

4. **Domain-Specific Methods**

- **Last Observation Carried Forward (LOCF):** In time series data, replace missing values with the last observed value. Suitable for temporal data.
- **Interpolation and Extrapolation:** Estimate missing values by interpolating or extrapolating based on surrounding data points. Useful for time series or spatial data.

3. Evaluate Imputation Methods

- **Compare Imputation Methods:** Assess the effectiveness of imputation methods by comparing model performance or validation metrics.
- **Sensitivity Analysis:** Perform sensitivity analysis to understand how different imputation methods impact the results.

4. Best Practices

- **Understand Missing Data Mechanism:** Choose imputation techniques based on the mechanism of missing data (MCAR, MAR, MNAR).

- **Document Decisions:** Keep detailed records of how missing data was handled for reproducibility and transparency.
- **Consider Multiple Techniques:** In some cases, using a combination of imputation techniques might be more effective.

12. What is A/B testing?

Answer: A/B testing, also known as split testing, is a method used to compare two versions of a variable to determine which one performs better in a given context. It is commonly used in marketing, web design, product development, and various other fields to optimize outcomes and make data-driven decisions.

Concepts of A/B Testing

1. **Objective:**
   - The main goal is to identify which of the two variants (A or B) yields a better outcome based on a specific metric or goal. This helps in making informed decisions to improve performance.
2. **Design:**
   - **Variant A (Control):** The original version of the element being tested. It serves as the baseline for comparison.
   - **Variant B (Treatment):** The modified version with changes intended to improve performance or achieve a different outcome.
3. **Experiment Setup:**
   - **Random Assignment:** Users or subjects are randomly assigned to either variant A or variant B to ensure that the comparison is fair and that the results are not biased by external factors.
   - **Sample Size:** A sufficient number of participants or data points are needed to ensure that the results are statistically significant.
4. **Metrics:**
   - **Primary Metric:** The main outcome measure that determines the success of the test. For example, this could be conversion rate, click-through rate, revenue, or user engagement.
   - **Secondary Metrics:** Additional measures that provide further insights but are not the primary focus of the test.
5. **Analysis:**
   - **Statistical Significance:** Statistical tests (such as t-tests or chi-squared tests) are used to determine if the differences observed between A and B are statistically significant and not due to random chance.
   - **Effect Size:** Measures the magnitude of the difference between A and B to understand its practical significance.

13. Is mean imputation of missing data acceptable practice?

Answer: Mean imputation, where missing values in a dataset are replaced with the mean of the observed values for that variable, is a common and straightforward method for handling missing data. However, its appropriateness depends on the context and specific characteristics of the data.

While mean imputation is a widely used and simple method, it is often not the best approach for handling missing data, especially when the proportion of missing values is large or when more sophisticated analyses are required. It is crucial to consider the nature of the missing data and the specific context of the analysis when choosing an imputation method.

14. What is linear regression in statistics?

Answer: Linear regression is a fundamental statistical method used to model and analyse the relationship between a dependent variable and one or more independent variables. The goal is to find the best-fitting line (or hyperplane, in the case of multiple independent variables) that describes how changes in the independent variables are associated with changes in the dependent variable.

Linear regression is a powerful and widely-used technique in statistics and data analysis that provides insights into the relationships between variables and enables predictions. Understanding its assumptions, limitations, and evaluation metrics is crucial for applying it effectively and interpreting results accurately.

15. What are the various branches of statistics?

Answer: Statistics is a broad field that encompasses various branches, each focusing on different aspects of data analysis, interpretation, and application. Here are the main branches of statistics:

1. Descriptive Statistics: Summarize and describe the main features of a dataset.

2. Inferential Statistics: Make predictions or inferences about a population based on a sample.

3. Probability Theory: Study the mathematical foundation of randomness and uncertainty.

4. Bayesian Statistics: Incorporate prior knowledge or beliefs along with data to make statistical inferences.

5. Multivariate Statistics: Analyse and interpret data involving multiple variables simultaneously.

6. Time Series Analysis: Analyse data collected over time to identify trends, seasonal patterns, and forecasting future values.

7. Experimental Design: Plan and conduct experiments to ensure valid and reliable results.

8. Non-Parametric Statistics: Analyse data without assuming a specific distributional form.

9. Statistical Computing: Develop and use computational tools for statistical analysis.

10. Econometrics: Apply statistical methods to economic data for empirical analysis.

11. Biostatistics: Apply statistical methods to biological and medical research.