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

Ans – True

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 - 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 - 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 - 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 – **Poisson**

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

a) True

b) False

Ans - 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

Ans – Hypothesis

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

Ans – 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

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

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.

Shape: The normal distribution appears as a bell curve when graphed. It is symmetric about the mean (average) and extends infinitely in both directions.

Probability Density Function (PDF): The general form of its probability density function is:

f(x)=2πσ21⋅e−2σ2(x−μ)2

- Here, (x) represents the random variable.
- (\mu) is the mean (also the median and mode).
- (\sigma) is the standard deviation (a measure of spread).

Properties:

- Most data in nature (e.g., heights of people, blood pressure) follows this distribution.
- The area under the normal distribution curve represents probability, and the total area under the curve sums to one.

Central Limit Theorem (CLT): The normal distribution is crucial because of the CLT. It states that the average of many samples from any distribution (with finite mean and variance) tends toward a normal distribution. This makes it applicable to various real-world scenarios.

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

Ans - Handling missing data is crucial for accurate analyses. Let's explore some common techniques for imputing missing values:

1. **Mean/Median/Mode Imputation**:
    - Replace missing values with the **mean** (for continuous data), **median** (when outliers are present), or **mode** (for categorical data).
    - Simple but can reduce variance.
2. **Linear Regression Imputation**:
    - Use a regression model to predict and replace missing values based on other variables.
    - Requires relationships between variables.
3. **Multiple Imputation**:
    - Generate multiple datasets with imputed values using statistical models.
    - Combine results to account for uncertainty.

12. What is A/B testing?

Ans - **A/B testing**, also known as **split testing**, is a research method used in marketing, web development, and user experience (UX) design. Here is how it works:

1. **Purpose**: A/B testing compares **two versions** of a web page, email, or other digital asset to determine which one performs better.
2. **Process**:
    - **Split Audience**: Marketers divide their audience into two groups.
    - **Variants**: They create **multiple versions** (A and B) of a specific variable (e.g., a webpage design, email subject line, or product feature).
    - **Comparison**: Users in each group see one of the variants.

    - **Effectiveness**: By analyzing user behavior (e.g., clicks, conversions), they determine which variant is more effective.
3. **Applications**:
    - **Emails**: Test different subject lines or content.
    - **Web Pages**: Compare layouts, colors, or copy.
    - **Apps**: Optimize features and user flows.

A/B testing helps organizations make data-driven decisions, improve user experiences, and enhance performance.

13. Is mean imputation of missing data acceptable practice?

Ans - **Mean imputation** is a common method for handling missing data, but it has both advantages and limitations.

1. **Advantages**:
    - **Simple**: Mean imputation is straightforward to implement.
    - **Preserves Sample Size**: It doesn't reduce the sample size.
    - **Maintains Central Tendency**: The mean reflects the central tendency of the data
2. **Limitations**:
    - **Bias**: Mean imputation assumes that missing values are missing at random (MAR). If they are not MAR, it can introduce bias.
    - **Underestimation of Variability**: Imputing means can underestimate the variability in the data.
    - **Impact on Relationships**: Mean imputation can distort relationships between variables.
3. **Alternatives**:
    - **Multiple Imputation**: Generates multiple imputed datasets to account for uncertainty.
    - **Regression Imputation**: Predict missing values using regression models.
    - **Domain-Specific Methods**: Use context-specific imputation techniques.

14. What is linear regression in statistics?

Ans - Linear regression is a statistical model that estimates the linear relationship between a scalar response (dependent variable) and one or more explanatory variables (independent variables)

1. **Purpose**: Linear regression helps us understand how the dependent variable changes as the independent variable(s) change.

Simple Linear Regression:

- In its simplest form, linear regression involves **two variables**:
    - **(y)** (dependent variable)
    - **(x)** (independent variable)
- The goal is to find a **straight line** (linear equation) that best represents the trend in the data
- **Formula**:
    - The linear regression equation is typically expressed as: [y = \beta_0 + \beta_1 x + \varepsilon]
        - (\beta_0) and (\beta_1) are the **regression coefficients**.
        - (\varepsilon) represents the **error term** (residuals).

Assumptions:

- **Homoscedasticity**: The error variance remains constant across different values of the independent variable.

- **Independence of Observations**: Data collected using valid sampling methods, with no hidden relationships among observations.
- **Normality**: The error terms follow a normal distribution.
- **Linearity**: The relationship between (x) and (y) is linear

Applications:

Predictive modeling (e.g., predicting house prices based on square footage).

Understanding cause-and-effect relationships (e.g., how advertising spending affects sales).

15. What are the various branches of statistics.

Ans - Statistics, as a field of study, encompasses various branches. Let's explore the two main branches:

1. **Descriptive Statistics**:
    - **Purpose**: Descriptive statistics involves organizing, summarizing, and displaying data.
    - **What It Does**:
        - Collects and organizes data.
        - Computes measures of central tendency (e.g., mean, median, mode) and variability (e.g., variance, standard deviation).
        - Presents data using tables, charts, and graphs.
        - **Example**:
            1. Average scores of college students in a math test.
            2. Average age of people who voted for a winning candidate.
            3. Length of a statistics book.

Inferential Statistics:

- **Purpose**: Inferential statistics draws conclusions or predictions about a **population** based on a **sample**.
- **What It Does**:
    1. Uses sample data to make inferences about larger populations.
    2. Performs hypothesis testing, regression analysis, and confidence intervals.

Example:

- Estimating the average income of all citizens based on a survey sample.
- Testing whether a new drug is effective using clinical trial data.