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: a) 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: 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: The square of a standard normal random variable follows what is called chi-squareddistribution

5. _______random variables are used to model rates.
   a) Empirical
   b) Binomial
   c) Poisson
   d) All of the mentioned

Ans: c) Poisson

6. 10. Usually replacing the standard error by its estimated value does change the CLT.
   a) True
   b) False

   Ans: 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

Ans: b) 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: 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: Outliers cannot conform to the regression relationship

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

Q10. What do you understand by the term Normal Distribution?

Ans: It’s referred to as the Gaussian distribution, is a symmetrical probability distribution that is characterized by a bell-shaped curve when graphed. In this distribution:

1. It’s symmetrical around the mean, which is located at the center of the curve.
2. The curve is bell-shaped, with tails extending infinitely in both directions but approaching, without touching, the horizontal axis.
3. The mean, median, and mode of the distribution are all equal and located at the center of the curve.
4. The total area under the curve is always equal to 1, representing the probability of all possible outcomes.
5. Many natural phenomena and measurements in various fields, such as heights, test scores, and errors in measurements, often follow a normal distribution.

It’s a fundamental concept in statistics and probability theory, extensively used due to its well-defined properties and its relevance in modelling various real-world phenomena.

Q11. How do you handle missing data? What imputation techniques do you recommend?

Ans Handling missing data is crucial in data analysis to avoid bias and ensure accurate results. There are various techniques to handle missing data:

1. Dropping missing values: This involves removing rows or columns with missing data. It's suitable when missing data is limited and doesn't significantly impact the analysis.
2. Mean/Median/Mode imputation: Replace missing values with the mean, median, or mode of the available data for that variable. It's a simple method but may not capture the true variability of the data.
3. Forward/Backward Fill: Propagate the last known value forward or backward to fill missing data in a time series.
4. Interpolation: Estimate missing values based on the relationships between other variables or neighboring data points.
5. Multiple Imputation: Use advanced statistical methods to create multiple imputed datasets by replacing missing values with plausible values multiple times to account for uncertainty.

Q12.What is A/B testing?

Ans: A/B testing is a method used in marketing, user experience design, and statistical analysis to compare two different versions of something

to determine which one performs better in achieving a specific goal or outcome. It involves dividing users or participants into two groups randomly - one group experiences the original or control version (A), while the other experiences a modified version (B). By measuring how each group responds or behaves, such as click-through rates, conversions, or user engagement, analysts can statistically assess which version is more effective and make informed decisions based on data to optimize for better performance. It helps in understanding user preferences, improving products, and optimizing strategies by identifying the version that yields better results.

Q13.Is mean imputation of missing data acceptable practice?

Ans: Mean imputation, which involves replacing missing values with the mean of the observed data, is a simple and commonly used method to handle missing data. However, while it's straightforward, it can introduce bias and distort the original distribution of the data, especially if the missing data is not missing at random. Despite its simplicity, mean imputation may not accurately represent the true variability within the dataset. As such, while it's widely used due to its simplicity, caution should be exercised when using mean imputation, and alternative methods like multiple imputation or predictive modeling should be considered to handle missing data more effectively without introducing significant bias.

Q14.What is linear regression in statistics?

Ans: Linear regression in statistics is a fundamental method used to model the relationship between a dependent variable and one or more independent variables by fitting a linear equation to observed data. It aims to find the best-fitting straight line (or hyperplane in multiple dimensions) that represents the relationship between the variables. This method helps in understanding and predicting the behavior of the dependent variable based on the independent variables. Essentially, linear regression analyzes how changes in the independent variables are associated with changes in the dependent variable, providing insights into patterns and making predictions.

Q15.What are the various branches of statistics?

Ans: The various Branches of statistics encompass various specialized areas within the field. Some main branches include:

1. Descriptive Statistics: Involves organizing, summarizing, and describing data using measures like mean, median, and mode.
2. Inferential Statistics: Focuses on making predictions or inferences about populations based on sample data.
3. Probability Theory: Deals with the likelihood of events occurring and forms the foundation of statistical reasoning.
4. Applied Statistics: Utilizes statistical methods to solve real-world problems in diverse fields like economics, medicine, and social sciences.
5. Biostatistics: Applies statistical methods to biological and health-related data analysis.
6. Econometrics: Applies statistical methods to economic data for modeling and forecasting economic relationships.
7. Data Science: Involves extracting insights and knowledge from data using statistical methods, machine learning, and computational techniques.

These branches collectively contribute to understanding data, making informed decisions, and drawing conclusions from observations in various fields.