FLIP ROBO

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: 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: d) All of the mentioned

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

Q10 and Q15 are subjective answer type questions, Answer them in your own words briefly.

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

Ans: Normal Distribution is a statistical probability distribution that is symmetric and bell-shaped. It is also known as the Gaussian distribution. In a normal distribution, the data tends to cluster around the mean, with the majority of values falling near the center and fewer values further away from the mean. The distribution is characterized by its mean and standard deviation, which determine its shape and spread. Many natural phenomena in fields such as social sciences, physical sciences, and economics can be approximated by a normal distribution

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

Ans: Commonly used techniques for handling missing data

1. Listwise deletion: This approach involves removing entire cases (rows) with missing data.

2. Pairwise deletion: This technique uses the available data for each analysis separately, excluding missing values only for the variables involved in that particular analysis. It allows retaining more data but can result in inconsistent sample sizes across analyses.

3. Mean or median imputation: Missing values are replaced with the mean or median value of the variable. This method is simple to implement but can distort the variable's distribution and may underestimate the variability.

4. Regression imputation: Missing values are imputed by predicting them from other variables using regression analysis. It takes into account the relationships among variables, but the imputed values may not be as precise as actual data.

5. Multiple imputation: This technique creates multiple plausible imputed datasets, considering the uncertainty associated with missing values. The analysis is performed on each imputed dataset, and the

results are combined to obtain valid inferences. Multiple imputation is generally preferred over single imputation methods because it accounts for the uncertainty due to missing data.

Imputation technique depends on various factors, such as the nature and extent of missingness, the variables involved, the underlying data distribution, and the research objectives

12. What is A/B testing?

Ans: A/B testing, also known as split testing, is a statistical technique used to compare two or more variations of a webpage, marketing campaign, or other elements to determine which one performs better in terms of a desired outcome.

The process involves dividing a sample of users or customers into two or more groups. One group (the control group) is exposed to the original or existing version (A), while the other groups (the test groups) are exposed to different variations (B, C, etc.) that include specific changes or modifications. The groups are randomly assigned to ensure unbiased results.

13. Is mean imputation of missing data acceptable practice?

Ans: Mean imputation of missing data is a commonly used technique due to its simplicity, but it has several limitations and potential drawbacks. While it may be acceptable in certain situations, it is generally advisable to use more sophisticated imputation methods that preserve the uncertainty associated with missing values.

14. What is linear regression in statistics?

Ans: Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables. It assumes a linear relationship between the variables and aims to find the best-fit line that minimizes the difference between the observed data points and the predicted values.

15. What are the various branches of statistics?

Ans: Statistics, as a field of study, encompasses several branches or subfields that focus on different aspects of data analysis and interpretation. Here are some prominent branches of statistics:

1. Descriptive Statistics
2. Inferential Statistics
3. Probability Theory
4. Biostatistics
5. Econometrics
6. Multivariate Analysis
7. Time Series Analysis
8. Spatial Statistics
9. Nonparametric Statistics
10. Statistical Learning/Machine Learning

These branches of statistics overlap and complement each other, offering a rich toolbox of methods and techniques for analysing and interpreting data in different domains.