WORKSHEET

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

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?

c) Centroid Limit Theorem

3. Which of the following is incorrect with respect to use of Poisson distribution?

c) Modeling contingency tables

4.

Point out the correct statement.

c) The square of a standard normal random variable follows what is called chi-squared distribution

5. \_\_\_\_\_\_ random variables are used to model rates.

c) Poisson

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

b) False

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

b) Hypothesis

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

9. Which of the following statement is incorrect with respect to outliers?

c) Outliers cannot conform to the regression relationship

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?

The normal distribution is also known as probability bell curve. It is symmetric about the mean and indicates

that values near the mean occur more frequently than the values that are farther away from the mean. This is used to

reprent large number of variables.

Important feature of normal distribution are:

Symmetry: The normal distribution is symmetric around its mean. left side of It means that the distribution mirrors the right side.

In a normal distribution, the mean, median, and mode are all equal and located at the center of the distribution.

Bell-shaped Curve: which indicate that most of the observations cluster around the central peak, and the probabilities for

values further away from the mean taper off equally in both directions.

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

Handling missing data is very important in data analysis, as it can significantly impact the results of your analysis and models.

Here are some common strategies and techniques for handling missing data, along with imputation methods:

Understand the Missing Data and its impact

Imputation Techniques: When we want to retain as much data as possible, these method can be used to fill in missing values:

Mean/Median/Mode Imputation

K-Nearest Neighbors (KNN) Imputation:

Regression imputation

Interpolation

Multipulation

12. What is A/B testing?

A/B testing : is a statistical method used to compare two versions of a webpage, product, or marketing strategy

to determine which one performs better. This widely used in digital marketing, product development, and user experience optimization.

13. Is mean imputation of missing data acceptable practice?

Mean imputation is a common technique for handling missing data, where missing values for a variable are replaced with the mean of the observed values for that

variable. Advantages and disadvantages are:

Advantages:

Retains Sample Size: By imputing missing values, we retain all available cases in dataset which is important when the sample size is small.

Quick Computation: It requires minimal computational resources compared to more complex imputation methods, which make it efficient for large datasets.

Disadvantages:

Reduces Variability: Mean imputation can distort the variability of the dataset. which can make estimates bias.

Ignores Relationships: Mean imputation does not take into account potential relationships between variables.

Impact on Correlations: It can inflate the correlation between variables, leading to incorrect conclusions in multivariate analyses.

Introduces Bias: mean imputation can introduce bias into the dataset, impacting the results of analyses and predictions.

14. What is linear regression in statistics?

This method used to model and analyze the relationship between one dependent variable and one or more independent variables.

It aims to establish a linear relationship between the dependent variable and the independent variables by fitting a linear equation to the observed data.

Types of Linear Regression:

Simple Linear Regression

Multiple Linear Regression

Assumptions of Linear Regression:

Linearity

Independence

Homoscedasticity

Normality

Estimation of Coefficients

Evaluation of the Model

R-squared

Adjusted R square

Mean Squared Error

F-statistic

Applications of Linear Regression:

Risk Assessment

Predactive analysis

Trend analysis

15. What are the various branches of statistics?

There are many branches of Statistics few are:

1. Descriptive Statistics

2. Inferential Statistics

3. Bayesian Statistics

4 Predactive analysis

5 Suvey Analysis