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?

a) Central Limit Theorem

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

b) Modeling bounded count data

4. Point out the correct statement.

d) All of the mentioned

5. ______ random variables are used to model rates.

c) Poisson

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

b) False

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

b) Hypothesis

8. 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?

- A probability distribution that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean.
- Normal distribution, also known as the Gaussian distribution will appear as a bell curve.
- Normal distributions are symmetrical, but not all symmetrical distributions are normal.

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

Filling the missing data with the mean or median value if it's a numerical variable. Filling the missing data with the mode (most frequent) string or character if it's a str type data. This is the basic/crude method for handling missing data.

- Deleting the Missing values (recommended: if NaN values are very less and dataset is very large)
- Deleting the Entire Row
- Deleting the Entire Column
- Imputing the Missing Value
- Replacing with Arbitrary Value
- Replacing with Mean
- Replacing with Mode
- Replacing with Median
- Replacing with Previous Value – Forward Fill
- Replacing with Next Value – Backward Fill
- Interpolation
- Impute the Most Frequent Value
- Imputation of Missing Values using sci-kit learn library
- Univariate Approach
- Multivariate Approach
- Nearest Neighbors Imputations (KNNImputer)

12. What is A/B testing?

A/B testing (also known as split testing or bucket testing) is a method of comparing two versions of a webpage or app against each other to determine which one performs better. A/B testing is essentially an experiment where two or more variants of a page are shown to users at random and statistical analysis is used to determine which variation performs better for a given conversion goal.

13. Is mean imputation of missing data acceptable practice?

It’s a popular solution to missing data, despite its drawbacks. Mainly because it’s easy. It can be really painful to lose a large part of the sample. But that doesn’t make it a good solution, and it may not help you find relationships with strong parameter estimates. Even if they exist in the population. On the other hand, there are many alternatives to mean imputation that provide much more accurate estimates and standard errors.

Two major drawbacks:

- Mean imputation does not preserve the relationships among variables
- Mean Imputation Leads to An Underestimate of Standard Errors

14. What is linear regression in statistics?

Linear regression analysis is used to predict the value of a variable based on the value of another variable. The variable you want to predict is called the dependent variable. The variable you are using to predict the other variable's value is called the independent variable.

15. What are the various branches of statistics?

- Data collection
- Descriptive statistics
- Inferential statistics