STATISTICS WORKSHEET-1 (ANSWERS)

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? 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

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? d) None of the mentioned

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, also known as the Gaussian distribution or bell-shaped distribution, is a continuous probability distribution that is widely used in statistics, probability theory, and data analysis. It is a symmetric probability distribution that is characterized by its mean (μ) and standard deviation (σ), and it is fully defined by these two parameters.

The probability density function (PDF) of the normal distribution is given by the formula:

f(x) = (1 / (σ * √(2π))) * e^(-((x - μ)^2) / (2 * σ^2))

where:

- x is the random variable representing the value of the variable being modeled
- μ is the mean of the distribution, which represents the center of the distribution
- σ is the standard deviation of the distribution, which represents the spread or variability of the data
- e is Euler's number, a mathematical constant approximately equal to 2.71828

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

a) Handling missing data is an important step in data analysis and modeling, as missing data can introduce bias and affect the validity of the results. There are several approaches to handling missing data, and the choice of imputation technique depends on various factors such as the type and pattern of missing data, the nature of the data, the purpose of the analysis, and the assumptions made about the missingness.

b) Here are some common techniques for handling missing data:

1. Listwise or complete case analysis.
2. Pairwise deletion.
3. Mean, Median and Mode Imputation.
4. Regression Imputation.
5. Multiple Imputation.
6. Model – based Imputation.

12. What is A/B testing?

A/B testing, also known as split testing or bucket testing, is a method used in statistics and marketing to compare two or more versions of a web page, advertisement, or other digital content to determine which version performs better in terms of a specific outcome or metric. A/B testing involves randomly dividing a sample or audience into multiple groups, each exposed to a different version of the content, and then measuring the performance of each version to determine which one yields the best results.

The basic steps of A/B testing typically include the following:

a) Formulate a hypothesis.
b) Design the experiment.
c) Collect Data
d) Analyze the Data
e) Interpret the results.

13. Is mean imputation of missing data acceptable practice?

YES.

14. What is linear regression in statistics?

Linear regression is a statistical method used to model the relationship between two or more variables, where one variable, called the dependent variable or response variable, is predicted based on one or more independent variables or predictor variables. Linear regression seeks to find the best-fitting straight line (or hyperplane, in the case of multiple independent variables) that describes the linear relationship between the variables.

The general equation for a simple linear regression model with one dependent variable (Y) and one independent variable (X) is:

$Y = \beta 0 + \beta 1X + \varepsilon$

where:

- Y is the dependent variable (the variable being predicted)
- X is the independent variable (the predictor variable)
- β0 is the intercept, which represents the value of Y when X is equal to 0
- β1 is the slope, which represents the change in Y for each unit change in X
- ε is the error term, which accounts for variability in Y that cannot be explained by the linear relationship with X.

15. What are the various branches of statistics?

The following are the various branches of statistics:

a) Descriptive statistics.
b) Inferential statistics.
c) Probability Theory.
d) Biostatistics.
e) Econometrics.
f) Social statistics.
g) Multivariate statistics.
h) Time series Analysis.
i) Statistical Computing.