ASSIGNMENT

STATISTICS

Worksheet :-1

1. Bernoulli random variables take (only) the values 1 and 0.

a) True b) False

Answer:- False

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?

Answer:- a) Central Limit Theorem

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

Answer:- b) Modelling 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

Answer:- d) All of the mentioned

5. _____ random variables are used to model rates.

Answers:- c) Poisson

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

Answer:- b) False

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

Answer:- b) Hypothesis

8. Normalized data are cantered at_____and have units equal to standard deviations of the original data

Answer:- a) 0

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

Answer:- c) Outliers cannot conform to the regression relationship

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

Answer:- Normal distribution, also known as the Gaussian distribution, is 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. In a normal distribution the mean is zero and the standard deviation is 1. It has zero skew and a kurtosis of 3. Normal distributions are symmetrical, but not all symmetrical distributions are normal.

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

Answer:- Missing data is a huge problem for data analysis because it distorts findings. It's difficult to be fully confident in the insights when you know that some entries are missing values. There are three types of missing data. These are Missing Completely at Random (MCAR) – when data is completely missing at random across the dataset with no discernible pattern. There is also Missing At Random (MAR) – when data is not missing randomly, but only within sub-samples of data. Finally, there is Not Missing at Random (NMAR), when there is a noticeable trend in the way data is missing.
It can be handled by using techniques like: -

- Using deletion methods to eliminate missing data
  The deletion methods only work for certain datasets where participants have missing fields. There are several deleting methods – two common ones include Listwise Deletion and Pairwise Deletion. It means deleting any participants or data entries with missing values. This method is particularly advantageous to samples where there is a large volume of data because values can be deleted without significantly distorting readings.

- Use regression analysis to systematically eliminate data

  Regression is useful for handling missing data because it can be used to predict the null value using other information from the dataset. There are several methods of regression analysis, like Stochastic regression.

- Using data imputation techniques
  Two data imputation techniques to handle missing data: Average imputation and common-point imputation. Average imputation uses the average value of the responses from other data entries to fill out missing values. But by using this can artificially reduce the variability of the dataset. Common-point imputation, on the other hand, is used when the data scientists utilise the middle point or the most commonly chosen value.

12. What is A/B testing?

Answer:- An AB test is an example of statistical hypothesis testing, a process whereby a hypothesis is made about the relationship between two data sets and those data sets are then compared against each other to determine if there is a statistically significant relationship or not.

13. Is mean imputation of missing data acceptable practice?

Answer:- Mean imputation is typically considered terrible practice since it ignores feature correlation.

14. What is linear regression in statistics?

Answer:- Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data. One variable is considered to be an explanatory variable, and the other is considered to be a dependent variable.

15. What are the various branches of statistics?

There are three branches of statistics:

- Data collection,
- Descriptive statistics
- Inferential statistics.