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
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
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
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
5. _____random variables are used to model rates.
   a) Empirical
   b) Binomial
   c) Poisson
   d) All of the mentioned
6. 10. Usually replacing the standard error by its estimated value does change the CLT.
   a) True
   b) False
7. 1. Which of the following testing is concerned with making decisions using data?
   a) Probability
   b) Hypothesis
   c) Causal
   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
   b) 5
   c) 1
   d) 10
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

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

The normal distribution, also known as the Gaussian distribution.

The normal distribution is a continuous probability distribution that is symmetrical around its mean, 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. The mean, median, and mode are all equal.

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

The two primary methods to solve the missing data is: imputation or the removal of data.

1. Imputation: The imputation method develops reasonable guesses for missing data. It's most useful when the percentage of missing data is low. If the portion of missing data is too high, the results lack natural variation that could result in an effective model.

2. Removing Data: When dealing with data that is missing at random, related data can be deleted to reduce bias. Removing data may not be the best option if there are not enough observations to result in a reliable analysis. In some situations, observation of specific events or factors may be required.

12. What is A/B Testing?

A/B testing is a basic randomized control experiment. It is a way to compare the two versions of a variable to find out which performs better in a controlled environment.

13. Is mean imputation of missing data acceptable practice?

No

1. Mean imputation reduces the variance of the imputed variables.
2. Mean imputation shrinks standard errors, which invalidates most hypothesis tests and the calculation of confidence interval.
3. Mean imputation does not preserve relationships between variables such as correlations.

14. What is linear regression in statistics?

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?

Statistics is mainly divided into the following two categories.

1. Descriptive statistics

2. Inferential statistics

1. Descriptive Statistics

In the descriptive statistics, the data is described in a summarized way. The summarization is done from the sample of the population using different parameters like mean or standard deviation. Descriptive statistics are a way of using charts, graphs, and summary measures to organize, represent, and explain a set of data.

Data is typically arranged and displayed in tables or graphs summarizing details such as histograms, pie charts, bars or scatter plots.

Descriptive statistics are just descriptive and thus do not require generalization beyond the data collected.

2. Inferential Statistics

In the Inferential Statistics, we try to interpret the meaning of descriptive statistics. After the data has been collected, analyzed, and summarized we use Inferential Statistics to describe the meaning of the collected data.

Inferential Statistics use the probability principle to assess whether trends contained in the research sample can be generalized to the larger population from which the sample originally comes.

Inferential Statistics are intended to test hypotheses and investigate relationships between variables and can be used to make population predictions.

Inferential Statistics are used to draw conclusions and inferences, i.e., to make valid generalizations from samples.