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

FLIP ROBO

Q10and Q15 are subjective answer type questions, Answer them in your own words briefly.

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

Ans. Normal Distribution is also called as bell curve. It 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. The normal distribution is the most important probability distribution in statistics because it fits many natural phenomena. For example, heights, blood pressure, measurement error, and IQ scores follow the normal distribution. The normal distribution model is motivated by the Central Limit Theorem.

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

Ans. Below are three common ways to handle missing data:

1. Mean or Median Imputation. When data is missing at random, we can use list-wise or pair-wise deletion of the missing observations.
2. Multivariate Imputation by Chained Equations (MICE) MICE assumes that the missing data are Missing at Random (MAR). It imputes data on a variable-by-variable basis by specifying an imputation model per variable.
3. Random Forest-This is a non-parametric imputation method applicable to various variable types that works well with both data missing at random and not missing at random.

Imputation is a technique used for replacing the missing data with some substitute value to retain most of the data/information of the dataset.

Imputation Techniques: -
1. Mean or Median Imputation
2. Multivariate Imputation by Chained Equations (MICE)
3. Random Forest

FLIP ROBO

You could find missing/corrupted data in a dataset and either drop those rows or columns, or decide to replace them with another value. In Pandas, there are two very useful methods: isnull () and dropna() that will help you find columns of data with missing or corrupted data and drop those values. If you want to fill the invalid values with a placeholder value (for example, 0), you could use the fillna() method.

12. What is A/B testing?

Ans. A/B testing also known as split testing or bucket testing. it is a method of comparing two versions of a variants against each other to determine which one performs better. A/B testing is basically 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?

Ans. Mean imputation is the practice of replacing null values in a data set with the mean of the data.

It is a non-standard, it uses Random Forest. It is use to predict the missing data. It also can be used for both i.e., continuous as well as categorical data and so it makes advantageous over other imputations.

There are some limitations too: -
1. Mean imputation does not preserve the relationship among variables. It preserves the mean of observed data. If data is missing completely at random, the estimate of the mean remains unbiased.
2.Mean Imputation leads to an underestimate of standard errors.

14. What is linear regression in statistics?

Ans. 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. For example, a modeler might want to relate the weights of individuals to their heights using a linear regression model.

A linear regression line has an equation of the form Y = mx + c, where X is the explanatory variable and Y is the dependent variable. The slope of the line is m, and c is the intercept (the value of y when x = 0).

Types of linear regression:
1. Simple linear regression
2. Multiple linear regressions
3.Logistic regression
4. Ordinal regression
5. Multinomial regression

15. What are the various branches of statistics?
Ans. Various branches of statistics are given below:

FLIP ROBO

1. Descriptive Methods: - This type of method consists of all the preliminary steps to final analysis and interpretation. As such this method includes the method of collection, methods of tabulation, measures of central tendency, measures of dispersion, measures of skewness, and analysis of time series. These methods bring out the various characteristics of data and help in summarizing and interpreting the salient features of the data. This method is also otherwise called descriptive statistics.
2. Analytical Methods: - This type of method consists of all those methods which help in the matter of analysis and comparison between any two or more variables. This includes the methods of correlation, regression analysis, association of attributes and the like. This method is also otherwise called analytical statistics.
3. Inductive Methods: - This type of method consists of all those procedures that help in the generalization or estimation over a phenomenon on the basis of random observation or partial data. This includes the procedure of interpolation, extrapolation, theory of probability and the like. This method is also otherwise called inductive statistics.
4. Inferential Methods: - This type of method consists of those procedures which help which in drawing inferences about the characteristics of the population on the basis of samples. As such, this method includes the theory of sampling, different tests of significance, statistical control etc. This method is also otherwise called inferential statistics.
5. Applied Methods: - This type of method consists of those procedures which are applied to the problems of real life. This includes the method of statistical quality control, sample survey, linear programming, inventory control and the like.