STATISTICS WORKSHEET-1

1. Bernoulli random variables take (only) the values 1 and 0.
a) True
b) False

Answer: - 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
b) Central Mean Theorem
c) Centroid Limit Theorem
d) All of the mentioned

Answer: - a) Central Limit Theorem

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

Answer: - b) modeling bounded count data

4. Point out the correct statement.
a) The exponent of 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.
a) Empirical
b) Binomial

c) Poisson
d) All of the mentioned

Answer: - 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?
a) Probability
b) Hypothesis
c) Causal
d) None of the mentioned

Answer: - b) Hypothesis

8. Normalized data are centered at ______ and have units equal to standard deviations of the original data.
a) 0
b) 5
c) 1
d) 10

Answer: - a) 0

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

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

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

Answer: - Normal Distribution in graphical representation is a bell shape curve that implies the mean, median and mode to be converging at the same point. 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, measurement error, and IQ scores follow the normal distribution. The normal distribution model is motivated by the Central Limit Theorem. The peak of a normal distribution consists of the maximum data points while the bottom part reduces in terms of frequency of data.

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

Answer: - Missing data (or missing values) is defined as the data value that is not stored for a variable in the observation of interest. The problem of missing data is relatively mutual in almost all research and can have a noteworthy effect on the conclusions that can be drawn from the data.

Data can be missing in the following ways:-

- Missing Completely At Random (MCAR): When missing values are randomly distributed across all observations, then we consider the data to be missing completely at random.
- Missing At Random (MAR): The key difference between MCAR and MAR is that under MAR the data is not missing randomly across all observations, but is missing randomly only within sub-samples of data.
- Not Missing At Random (NMAR): When the missing data has a structure to it, we cannot treat it as missing at random.

Imputation Techniques: -

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

We 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 us to find columns of data with missing or corrupted data and drop those values. If we 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?

Answer: - A/B testing also known as split testing. An A/B 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 the replacement of a missing observation with the mean of the non-missing observations for that variable.
It is a non-standard method, it uses Random Forest. It is used 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. Any statistic that uses the imputed data will have a standard error that's too low.

14. What is linear regression in statistics?

Answer: - Linear regression is a basic and commonly used type of predictive analysis. 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 has an equation of the form,

$$Y = mx + c,$$

Where, x is the explanatory variable and Y is the dependent variable, m = regression coefficient, 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?

Answer: - Various branches of statistics are given below: -

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**.**