STATISTICS_WORKSHEET-1

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

   Answer – 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?

   Answer – Central Limit Theorem

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

   Answer – All of the mentioned.

4. Point out the correct statement.

   Answer - All of the mentioned.

5. _____ random variables are used to model rates.

   Answer – Poisson

6. Usually replacing the standard error by its estimated value does change the CLT.

   Answer – False

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

   Answer – Hypothesis

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

   Answer – 0

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

   Answer - Outliers cannot conform to the regression relationship.

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

Answer – Normal Distribution is defined by the probability density function for a continuous random variable. It is defined as a function which is integrated in the range or interval (x to x+dx)

It is given by the formula $=f(x,\mu,\sigma)=\frac{1}{\sigma\sqrt{2\pi}}e^{\frac{-(x-\mu)^2}{2\sigma^2}}$

$$where\ \ x\ is\ variable,$$

$$\mu\ is\ mean,$$

$$and\ \sigma\ is\ standard\ deviation$$

Here mean is the summation of the variables divided by the number of variables and standard deviation is how far the data is spread.

Approximately 68% of data falls in one standard deviation of mean

95 % of data falls within two standard deviation of mean and 99.7 % falls within three standard deviation of the mean.

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

Answer – Missing data is the data having NaN value or null value in a dataset it can be present in a dataset and we can use various pandas function to operate the missing data.

First of all we have to detect missing data using isnull() and notnull() function in Data Frame.

Second step is cleaning and filling missing data using Pandas fillna() is used to fill non-null data

In a couple of ways.

Dropping missing values can also be done using dropna() function along with the axis argument.

By default axis=0, along rows dropping is done.

Five imputation techniques can be used

a. Hot-Deck Imputation- In this imputation technique missing data is replaced by a similar selected record.
b. Cold-Deck Imputation- In this method data is replaced from some other datasets.
c. Mean-Substitution- In this method missing data is replaced with the mean of that variable for all other cases.
d. Regression Imputation- In this imputation regression model is estimated to predict observed values of a variable based on other variables, and that model is then used to impute values in cases where the value of that variable is missing.

12. What is A/B testing?

Answer- A/B is a statistical hypothesis testing of two randomized variable A and B. A/B testing is a shorthand for a simple controlled experiment in which two samples (A and B) of a single vector-variable are compared.

13. Is mean imputation of missing data acceptable practice?

Answer - Imputation of mean data is not an acceptable practice since:

a. It does not preserve the relationships among variables
b. It leads to the underestimate of Standard Errors

14. What is linear regression in statistics?

Answer – Linear Regression tries to show the relationship between two variables by applying a linear equation to observed data. One variable is dependent and the other one is independent. In the linear regression equation correlation coefficient is used to measure the extent of relationship. Range of this coefficient varies from -1 to +1.

Linear regression equation is written in the form of $Y = a + bX$

Where X is independent variable and plotted against the x-axis.

Y is the dependent variable and plotted against the Y axis.

The slope of the line is b, and a is the intercept.

15. What are the various branches of statistics?

Answer – Two branches of statistics are present

First is Descriptive and another one is inferential statistics.

In the descriptive statistics the data is summarized through the given observations.

The summarization can be done using parameters like mean or standard deviation.

In the descriptive statistics representation of collection of data is done using tables, graphs and summary measures.

Descriptive statistics are also categorised into four different categories :

Measure of frequency, Measure of dispersion, Measure of central tendency, Measure of position.

In the inferential statistics is used to interpret the meaning of descriptive statistics. It means once the data has been collected, analysed, summarized then we can use the stats to describe the meaning of collected data. Various methods like Z-score test, Chi-square test and correlation coefficient are used in this statistics.