ANSWERSHEET FOR STATISTICS WORKSHEET - I

Date – 016/04/2023

1. Answer Option "A" True
2. Answer Option "A" Central Limit Theorm
3. Answer Option "Modeling Bound Count Data"
4. Answer Option "C" Chi Squared Distribution
5. Answer Option "C" Poisson
6. Answer Option "B" False
7. Answer Option "B" Hypothesis
8. Answer Option "A" 0
9. Answer Option "C"

10. **Normal Distribution** – The Normal Distribution also know as Gaussian Distribution, is used as a continuous probability distribution in Statistics to understand or show real-valued random variables whose distribution is not known.

11. **How to Handle Missing Data and Imputation techniques to use** – In case of Missing Data we have to check for missing data in the data frame which is genrally present in the form of Null Values, NaN, or Missing Values for this we use **isnull() or isna()** function in Pandas and find the missing values in the data frame, or same can be done by checking for summary Statistics like Mean, Median and Standard Deviation if these values differ significantly from your expectations there can be missing data, Data Visualization can also be used find missing values by plotting histograms, scatter plots or box plot. After finding the Null or the Missing values in the Data frame we can do following -
    Deletion – Deleting the missing data from Dataset (Listwise or Pairwise)
    Imputation – Estimating the missing values and filling them, common imputation techniques are –
    a. Mean Imputation – Replacing missing values with the mean of available data
    b. Median Imputation - Replacing missing values with the median of available data
    c. Mode Imputation - Replacing missing values with the mode of available data\
    d. Regressive Imputation – Predicting the missing values using Regression Model

    Following are commands used for Imputation –

    a. PANDAS – **fillna()**
    b. Scikit – Learn – Python library for ML
    c. KNNImputer – fill the missing values using the K-Nearest Neighbors Algorithm

12. **A/B Testing –** A/B Testing is a statistical method of comparing two versions (example – app/webpage) to identify which performs better so that it can be used in marketing, advertising, product development or enhance user experience. It is done by doing the following steps –

a. Define the Hypothesis
b. Design the test
c. Determine the Sample size
d. Run the test
e. Analyze the data
f. Draw Conclusions

13. **Mean Imputation of Missing Data** – Mean imputation is not a acceptable practice in all cases as it leads to **Bias, Loss of Information and Inaccurate estimates**. Though it is an easy and convenient practice but should be avoided, Mean imputation can only be a acceptable practice missing data is very low and distribution is very close to normal

14. **Linear Regression in Statistics** – Linear Regression is a Statistical Method which is mainly used when one has to find a relation between a dependent variable and one or more independent variables, the main goal to use the technique is to find a linear equation between dependent and independent variables(s).

15. **Branches of Statistics** – The Main branches in context of our learning are -

    a. Descriptive Statistics
    b. Inferential Statistics

Submitted by – Ashutosh Mishra