STATISTICS WORKSHEET 1

Q1 to Q9

Q1. Bernoulli random variables take (only) the values 1 and 0.
(A) True

Q2. Which of the following theorem states that the distribution of averages of aid variables, properly normalized, becomes that of a standard normal as the sample size increases?
(A) Central limit theorem

Q3. Which of the following is incorrect with respect to use of Poisson distribution?
(B) Modeling bounded count data

Q4. Point out the correct statement
(D) All of the mentioned

Q5. ______ random variables are used to model rates.
(C) Poisson

Q6. Usually replacing the standard error by its estimated value does change the CLT.
(B) False

Q7. Which of the following testing is concerned with making decisions using data?
(B) Hypothesis

Q8. 4. Normalized data are centred at______ and have units equal to standard deviations of the original data.
(A) 0

Q9. Which of the following statement is incorrect with respect to outliers.
(C) Outliers cannot conform to the regression relationship

Q10 to Q15

Q10. What do you understand by the term Normal Distribution?
Ans: - The Normal distribution is also known as **Gaussian** or **Gauss distribution**. The normal distribution is an important class of statistical distribution that applies in the most Machine Learning Algorithms.
Importance of the Normal Distribution:

- The statistical hypothesis test assumes that the data follows a normal distribution
- Both linear and non-linear regression assumes that the residual follows the normal distribution.
- Most statistical software programs support some of the probability functions for normal distribution as well

These are the parameters of the Normal Distribution:

- Mean
- Standard Deviation

Following are the properties of the Normal Distribution:

- It is symmetric
- The mean, median, and mode are equal
- Empirical Rule applies to all the Normal Distributions.

Q11. How do you handle missing data? What imputation techniques do you recommend?
Ans: - Missing data can have a huge impact on any project that's why we must find effective solutions to handle the missing data. We can handle the missing data by-

- Deletion- Listwise, Pairwise, Entire variables
- Imputation- mean, median and mode, K Nearest Neighbours (KNN), Linear regression, Last observation carried forward or Observation carried backward, Linear Interpolations etc

I recommend following imputation techniques for missing data-

- We can use Listwise deletion technique for small datasets.
- Fixed value imputation technique
- Mean imputation for numerical values
- Linear regression technique for missing value prediction for numerical values.
- KNN for nominal values

Q12. What is A/B testing?

Ans: -

- A/B testing is a basic randomized control experiment. With the help of A/B testing we can compare two versions of a variable to find out which version performs better compared to other in a controlled environment. It is also known as **split testing** or **bucket testing**.
- A data scientist collects and studies the data available to help optimize the website for a better consumer experience. And for this, it is essential to know how to use various statistical tools, especially the concept of A/B Testing. It is a hypothetical testing methodology for making decisions that estimate population parameters based on sample statistics.
- A/B testing allows individuals, teams and companies to make careful changes to their user experiences while collecting data on the results. This allows them to construct hypotheses and to learn why certain elements of their experiences impact user behaviour. In another way, they can be proven wrong—their opinion about the best experience for a given goal can be proven wrong through an A/B test.

Process of A/B testing-

- **Collect data to** Look for pages with low conversion rates or high drop-off rates that can be improved.
- **Identifying goals** are the metrics that you are using to determine whether or not the variation is more successful than the original version.
- **Generate hypothesis**
- Create Variations
- Run experiment
- Analyse the outcomes

Q13. Is mean imputation of missing data acceptable practice?

Ans: - Mean imputation of the missing data is a popular technique. But this technique has some serious disadvantages and that is why it is not considered as a good practice. Experts agrees that "mean imputation should be avoided when possible".

- Mean imputation reduces the variance of the imputed variables while increasing bias. Meaning, it changes the variance of the dataset, making it less accurate. Hence, the model will be less accurate and the confidence interval will also be narrower. As well as it ignores feature correlation.

Q14. What is linear regression in statistics?

Ans: **- Linear regression** is a regression model that estimates the relationship between one independent variable and one dependent variable using a straight line.

- Linear regression analysis is used to predict the value of a variable based on the value of another variable.
- This form of analysis estimates the coefficients of the linear equation, involving one or more independent variables that best predict the value of the dependent variable.
- Linear regression fits a straight line or surface that reduces the differences between predicted and actual output values.

Q15. What are the various branches of statistics?

Ans: **- Statistics** is concerned with developing and studying different methods for collecting, analysing and presenting the empirical data. There are two broad categories of Statistics: **Descriptive Statistics** and **Inferential Statistics**. These two categories together give us more powerful tool for description and prediction of data.

- **Descriptive Statistics**: It describes the important characteristics/ properties of the data using the measures the central tendency like **mean/ median/mode** and the measures of dispersion like **range**, **standard deviation**, **variance** etc.
- **Inferential Statistics**: The goal of the inferential statistics is to draw conclusions from a sample and generalize them. It determines the probability of the characteristics of the sample using probability theory. The most common methodologies used are **hypothesis tests**, **Analysis of variance**, etc.

---- -:- ---- -:- ---- -:- ----