# STATISTICS WORKSHEET-1

Submitted By: - CHETAN SHARMA (INTERN23)

Batch No: - 1836

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

Ans.:- 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?

Ans:- A (Central Limit Theorem)

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

Ans:- B (Modeling bounded count data)

4.) Point out the correct statement.

Ans:- D (All of the mentioned)

5.) _____ random variables are used to model rates.

Ans:- C (Poisson)

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

Ans:- B (False)

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

Ans:- B (Hypothesis)

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

Ans:- A (0)

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

Ans:- C (Outliers cannot conform to the regression relationship)

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

Ans:- The normal (or Gaussian) distribution is one particular kind of a bell shaped curve. It is unimodal (that is, there is one peak"), symmetric (that is you can flip it around its mid-point) and its mean, median and mode are all equal. However, it is only one such distribution - others meet all those conditions and are not normal. Many things are approximately normally distributed, for example the heights of adult human females or males, IQ, etc.

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

Ans: - According to data scientists, there are three types of missing data. These are Missing Completely at Random (MCAR) – when data is completely missing at random across the dataset with no discernable pattern. There is also Missing At Random (MAR) – when data is not missing randomly, but only within sub-samples of data. Finally, there is Not Missing at Random (NMAR), when there is a noticeable trend in the way data is missing.

**Best techniques to handle missing data:-**

**1. Complete Case Analysis(CCA):-** This is a quite straightforward method of handling the Missing Data, which directly removes the rows that have missing data i.e we consider only those rows where we have complete data i.e data is not missing.

- Easy to implement.
- No Data manipulation required.

**2. Arbitrary Value Imputation: -** This is an important technique used in Imputation as it can handle both the Numerical and Categorical variables. This technique states that we group the missing values in a column and assign them to a new value that is far away from the range of that column. Mostly we use values like 99999999 or -9999999 or “Missing” or “Not defined” for numerical & categorical variables.

- Easy to implement.
- We can use it in production.
- It retains the importance of “missing values” if it exists.

**3. Frequent Category Imputation: -** This technique says to replace the missing value with the variable with the highest frequency or in simple words replacing the values with the Mode of that column. This technique is also referred to as Mode Imputation.

- Implementation is easy.
- We can obtain a complete dataset in very little time.
- We can use this technique in the production model.

## 12.) What is A/B testing?

Ans:- A/B testing, or split testing as it's sometimes known, is the creation of at least one variant to test against a current webpage to ascertain which one performs better in terms of agreed metrics such as revenue per visitor (for e-commerce websites) or conversion rate. Split-testing is an important part of any conversion rate optimisation (CRO) programme and helps you to build a business case for making informed, evidence-led changes to your website. A/B testing is only part of an effective CRO programme. A/B testing is useful because it helps you to validate your hypotheses. These hypotheses are derived out of independent, objective research rather than guesses, hunches and well-meaning opinions.

## 13.) Is mean imputation of missing data acceptable practice?

Ans: - The process of replacing null values in a data collection with the data's mean is known as mean imputation. Mean imputation is typically considered terrible practice since it ignores feature correlation. Consider the following scenario: we have a table with age and fitness scores, and an eight-year-old has a missing fitness score. If we average the fitness scores of people between the ages of 15 and 80, the eighty-year-old will appear to have a significantly greater fitness level than he actually does. Second, mean imputation decreases the variance of our data while increasing bias. As a result of the reduced variance, the model is less accurate and the confidence interval is narrower. (I don't know much about this, I am in Learning Phase, and I'm working on it).

## 14.) What is linear regression in statistics?

Ans: - Linear regression is a basic and commonly used type of predictive analysis. A linear regression model in context of machine learning/statistics is basically a linear approach for modelling the relationships between the dependent variable (known as the result) and your independent variable(s) (known as 'features').

If our model has only one independent variable then it is called simple linear regression, else it is called multiple linear regressions.

So in essence in linear regression we try to model our dependent variable as the algebraic sum of some parameter times our independent variable(s) or their exponents or both. Each of the coefficients of independent variables is the 'weight' our model is assigning to that feature(s) and we train our model on these coefficients to better fit the data.

In machine learning and statistics when we talk about linear regression, we mean the model is linear in terms of parameters and not necessarily in terms of features - meaning that it's perfectly valid to have different and higher powers of same features, but their coefficient need to be different.

## 15.) What are the various branches of statistics?

Ans: - There are two main branches of statistics –

- Inferential Statistic.
- Descriptive Statistic.

**Inferential Statistics:** Inferential statistics used to make inference and describe about the population. These stats are more useful when it's not easy or possible to examine each member of the population.

**Descriptive Statistics:** Descriptive statistics are used to get a brief summary of data. You can have the summary of data in numerical or graphical form.