**STATISTICS WORKSHEET-1**

Ans 1. True

Ans 2. Central Limit Theorem

Ans 3. Modeling bounded count data

Ans 4. All of the mentioned

Ans 5. Poisson

Ans 6. False

Ans 7. Hypothesis

Ans 8. 0

Ans 9. Outliers cannot conform to the regression relationship

Ans 10. Normal distribution 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. The normal distribution is the most common type of distribution assumed in technical stock market analysis and in other types of statistical analyses.

Ans 11. There are some types of handling missing data. i.e.

### Listwise or case deletion

Listwise deletion is the most frequently used method in handling missing data, and thus has become the default option for analysis in most statistical software packages. Some researchers insist that it may introduce bias in the estimation of the parameters. However, if the assumption of MCAR is satisfied, a listwise deletion is known to produce unbiased estimates and conservative results. When the data do not fulfill the assumption of MCAR, listwise deletion may cause bias in the estimates of the parameters.

1. **Mean substitution**

In a mean substitution, the mean value of a variable is used in place of the missing data value for that same variable. This allows the researchers to utilize the collected data in an incomplete dataset. However, with missing values that are not strictly random, especially in the presence of a great inequality in the number of missing values for the different variables, the mean substitution method may lead to inconsistent bias.

### Pairwise deletion

Pairwise deletion is known to be less biased for the MCAR or MAR data, and the appropriate mechanisms are included as covariates. However, if there are many missing observations, the analysis will be deficient.

Imputation is a technique used for replacing the missing data with some substitute value to retain most of the data/information of the dataset.

Imputation techniques are:

### 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. This method is also popularly known as “Listwise deletion”.

### 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.

### 3. Frequent Category Imputation

Ans 12. A/B testing, also known as split testing, refers to a randomized experimentation process wherein two or more versions of a variable (web page, page element, etc.) are shown to different segments of website visitors at the same time to determine which version leaves the maximum impact and drives business metrics. The version that moves your business metric(s) in the positive direction is known as the ‘winner.’ Implementing the changes of this winning variation on your tested page(s) / element(s) can help optimize your website and increase business ROI.The metrics for conversion are unique to each website. For instance, in the case of eCommerce, it may be the sale of the products. Meanwhile, for B2B, it may be the generation of qualified leads.

Ans 13. 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.

Ans 14. Linear regression analysis is used to predict the value of a variable based on the value of another variable. The variable you want to predict is called the dependent variable. The variable you are using to predict the other variable's value is called the independent 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 minimizes the discrepancies between predicted and actual output values. There are simple linear regression calculators that use a “least squares” method to discover the best-fit line for a set of paired data. You then estimate the value of X (dependent variable) from Y (independent variable).

Ans 15. There are three real branches of statistics: data collection, descriptive statistics and inferential statistics.

1. **Data Collection**

Data collection is all about how the actual data is collected. For the most part, this needn’t concern us too much in terms of the mathematics (we just work with what we are given), but there are significant issues to consider when actually collecting data.

1. **Descriptive Statistics.**

Descriptive statistics is the part of statistics that deals with presenting the data we have. This can take two basic forms – presenting aspects of the data either visually (via graphs, charts, etc.) or numerically (via averages and so on).

1. **Inferential Statistics**

Inferential statistics is the aspect that deals with making conclusions about the data. This is quite a wide area; essentially you are asking ‘What is this data telling us, and what should we do?’

For example, a number of cars are driving too fast). Note, though, that this may not be the case; everyone might be driving at a perfectly acceptable speed, and the accidents are down to something other than speed (a blind spot or a pothole, for example). This is inferential statistics: take the data you have and make an ‘inference’ or ‘conclusion’ from it. We shall see much more of this later when we discuss things such as hypothesis testing, where we test to see whether the data supports a belief that we have.