STATISTICS WORKSHEET -1

ANS1- a-True

ANS2- a-Central Limit Theorem

ANS3- b-modeling bounded count data

ANS4- d-all of the mentioned

ANS5- Poisson

ANS6- b-false

ANS7- b-hypothesis

ANS8- a-0

ANS9- c-outliers cannot conform to the regression relationship

ANS10- NORMAL DISTRIBUTION:

- The normal distribution is the most significant probability distribution.

- In a normal distribution, data is symmetrically distributed with no skew. When plotted on a graph, the data follows a bell shape, with most values clustering around a central region and tapering off as they go further away from the center.

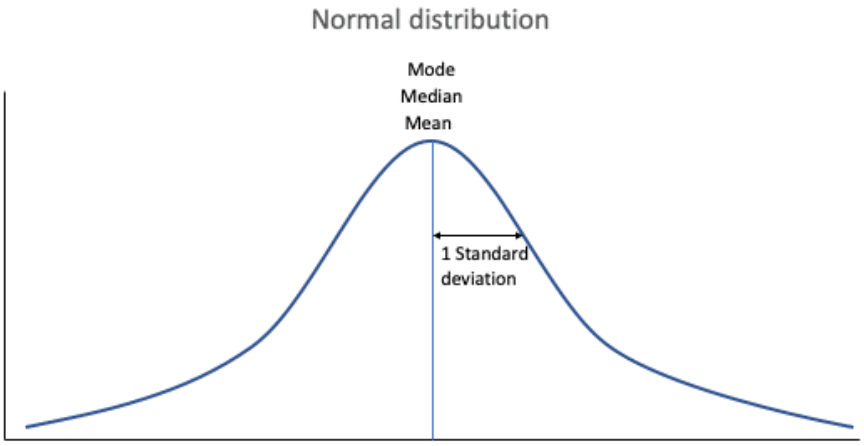
- Normal distributions are also called Gaussian distributions or bell curves because of their shape.

- Normal distributions have key characteristics that are easy to spot in graphs:

1-The mean, median and mode are exactly the same.

2-The distribution is symmetric about the mean—half the values fall below the mean and half above the mean.

3-The distribution can be described by two values: the mean and the standard deviation



ANS11- MISSING DATA:

Missing data, also known as missing values, is where some of the observations in a data set are blank. Missing data is a problem because it adds ambiguity to your analysis.

WHY A DATA GOES MISSING?

Before deciding which approach to employ, data scientists must understand why the data is missing.

1-Missing at Random (MAR)

Missing at Random means the data is missing relative to the observed data. It is not related to the specific missing values. The data is not missing across all observations but only within sub-samples of the data. It is not known if the data should be there; instead, it is missing given the observed data. The missing data can be predicted based on the complete observed data.

2-Missing Completely at Random (MCAR)

In the MCAR situation, the data is missing across all observations regardless of the expected value or other variables. Data scientists can compare two sets of data, one with missing observations and one without. Using a t-test, if there is no difference between the two data sets, the data is characterized as MCAR. Data may be missing due to test design, failure in the observations or failure in recording observations. This type of data is seen as MCAR because the reasons for its absence are external and not related to the value of the observation.

3-Missing not at a random(MNAR)

The MNAR category applies when the missing data has a structure to it. In other words, there appear to be reasons the data is missing. In a survey, perhaps a specific group of people – say women ages 45 to 55 – did not answer a question. Like MAR, the data cannot be determined by the observed data, because the missing information is unknown. Data scientists must model the missing data to develop an unbiased estimate. Simply removing observations with missing data could result in a model with bias.

HANDLING A MISSING DATA:

When dealing with missing data, data scientists can use two primary methods to solve the error: imputation or the removal of data.

1-The imputation method develops reasonable guesses for missing data. It's most useful when the percentage of missing data is low. If the portion of missing data is too high, the results lack natural variation that could result in an effective model.

2-The other option is to remove data. When dealing with data that is missing at random, related data can be deleted to reduce bias. Removing data may not be the best option if there are not enough observations to result in a reliable analysis. In some situations, observation of specific events or factors may be required.

IMPUTATION TECHNIQUE TO HANDLE MISSING DATA:

Input Data Validation – Discard Data Instance with Missing Data

Most trivial of all the missing data imputation techniques is discarding the data instances which do not have values present for all the features. In other words, before sending the data to the model, the consumer/caller program validates if data for all the features are present. If the data for all of the features are not present, the caller program do not invoke the model at all and takes on some value or show exceptions. For beginners, this could be a technique to start with. If this technique is used during training model training/testing phase, it could result in model bias.

ANS12- A/B TESTING:

A/B testing allows us to compare two versions of something to learn which is more effective. It removes the guess work from decision making and let the data decide the path forward. It includes application of statistical hypothesis testing, which is used in the field of statistics. It is a way to compare two versions of a single variant A and B of the variable, which then determine which variant is more effective.

ANS13- MEAN IMPUTATION:

In mean imputation technique, it preserves the mean of all observed data. So even if the data are missing completely at random, the estimated of the mean remain unbiased and unchanged. Also, by imputing the mean, we are able to keep the sample size up to full sample size.

ANS14- LINEAR REGRESSION:

Linear Regression is a statistical method which is used to understand a relationship between explanatory variable and a response variable. We assume that there is a linear relationship between the explanatory and response variable. If we add a unit to explanatory variable, then it will affect the response variable in the same way.

ANS15- BRANCHES OF STATISTICS:

Statistics is the branch of science that provides tools for decision making in the face of uncertainty.

Branches of statistics:

1- Descriptive Statistics- It helps in summarizing and organizing any data set characteristics. Helps in representing data in both classification and diagrammatic way.
   It has two parts:
   a- Central Tendency Measures- It includes the calculation of Mean, Median, Mode for better understanding.
   b- Variability Measures- It includes calculation of Ranges, Quartiles, Variance and Standard Deviation.
2- Inferential Statistics- It helps in finding the conclusion regarding the population after analysis of samples drawn from the population.
   Different types of Inferential Statistics:
   a- Regression Analysis
   b- Analysis of Variance
   c- Analysis of Covariance
   d- Statistics of Significance
   e- Correlation Analysis