STATISTICS WORKSHEET-1

1. Bernoulli random variable take (only) the values 0 and 1
   a) True

2. Which of the following theorem states that the distribution of averages of iid variables, properly normalised, becomes that of a standard normal as a sample size increases?
   a) Central Limit Theorem

3. Which of the following is incorrect with respect to use of poisson distribution?
   d) modelling bounded count data

4. Point out the correct statement

   d) all of the mentioned

5. _____ random variables are used to model rates.

   c) poisson

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

   a) True

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

   b) Hypothesis

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

   a) 0

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

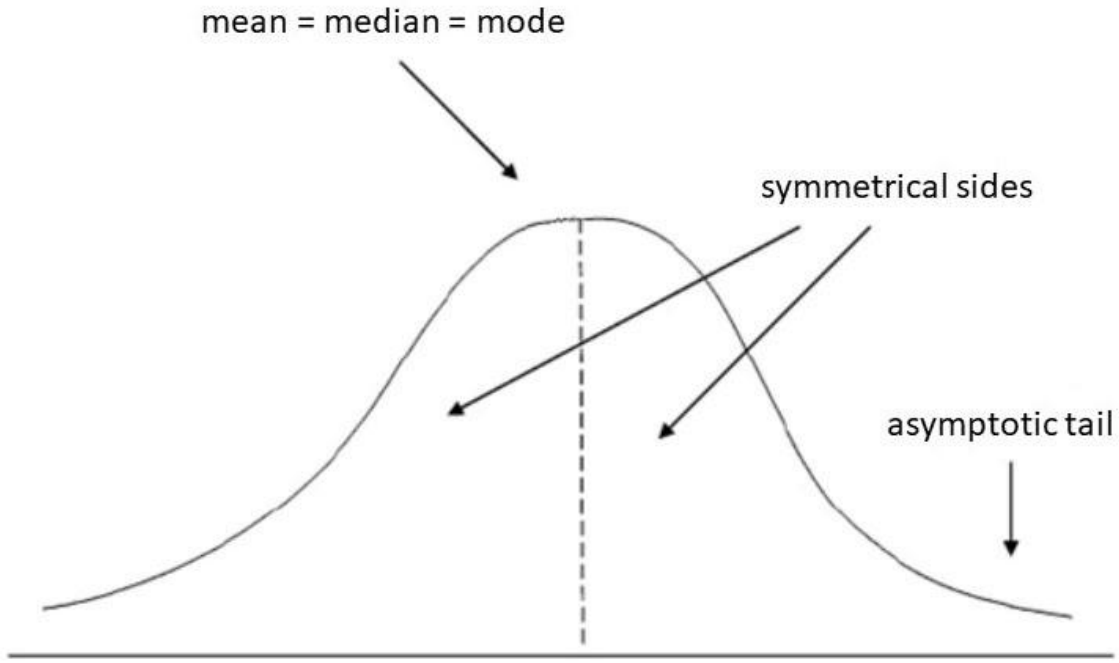
   c)Outliers cannot perform regression relationship.

10. What do you understand by the term normal distribution.

Normal distribution is also known as bell curve distribution is the most common distribution for independent randomly generated variables.

In this distribution two parameters are needed for the distribution graph.

1. Mean or average – which is the maximum of the graph and about which the graph is symmetric mean of nora=mal distribution is 0.
2. Standard deviation – which determines the amount of deapersion away from the mean. Maximum standard deviation for normal distribution is 3.



This is the graph of normal distribution in which mean, median, mode is equal to 0 and standard deviation is equal to -3 to 3which carries 99% od data and still if some data left than that data will count as outliers.

Outliers may affect the distribution graph and look like skewed data distribution whether its right skewed or left skewed data. it will simply effect the model building and model training testing and ultimately model accuracy so before going further we need to deal with outliers by different options that are given like z score , and many more.

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

Missing data is a huge problem to handle because if we can't handle missing data then we can't be confident about the insights of data that we created because of the missing values. Hence they must be addressed before insights.

Differenet types of missing data

1. MCAR (Missing Completely at random)—when data is completely missing at random across the dataset without any pattern.
2. MAR(Missing at random) – When data is missing not randomly but only sub samples of data.
3. NMAR(Not missing at random) – when there is noticeable trend the way data is missing.

Imputation Technique for missing data

1. A. Average Imputation—It will replace the missing data from the dataset to the average of that column and impute that average to the missing data place.

b. Common point imputation – It will take the common value that are repeated in the column(especially in median or mode) and replace it in the place of missing data.

2. Regression imputation

The result of regressing the missing variable on other factors to get a predicted value.As a result, instead of utilising the mean, you're relying on the anticipated value, which is influenced by other factors. This keeps the associations between the variables in the imputation model, but not the variability around the anticipated values.

3. Label Encoder – it will replace the nan values to the categorical form like 0,1 and many more.
4. Replacing missing values to the value of column which comes highest time in column so that the chances that missing values belongs to that is higher.
5. Deletion of missing values – if the missing values are high in the particular row then we simply delete it.

12. What is A/B testing?

A/B testing (also known as split testing or bucket testing) is a method of comparing two versions of a webpage or app against each other to determine which one performs better.

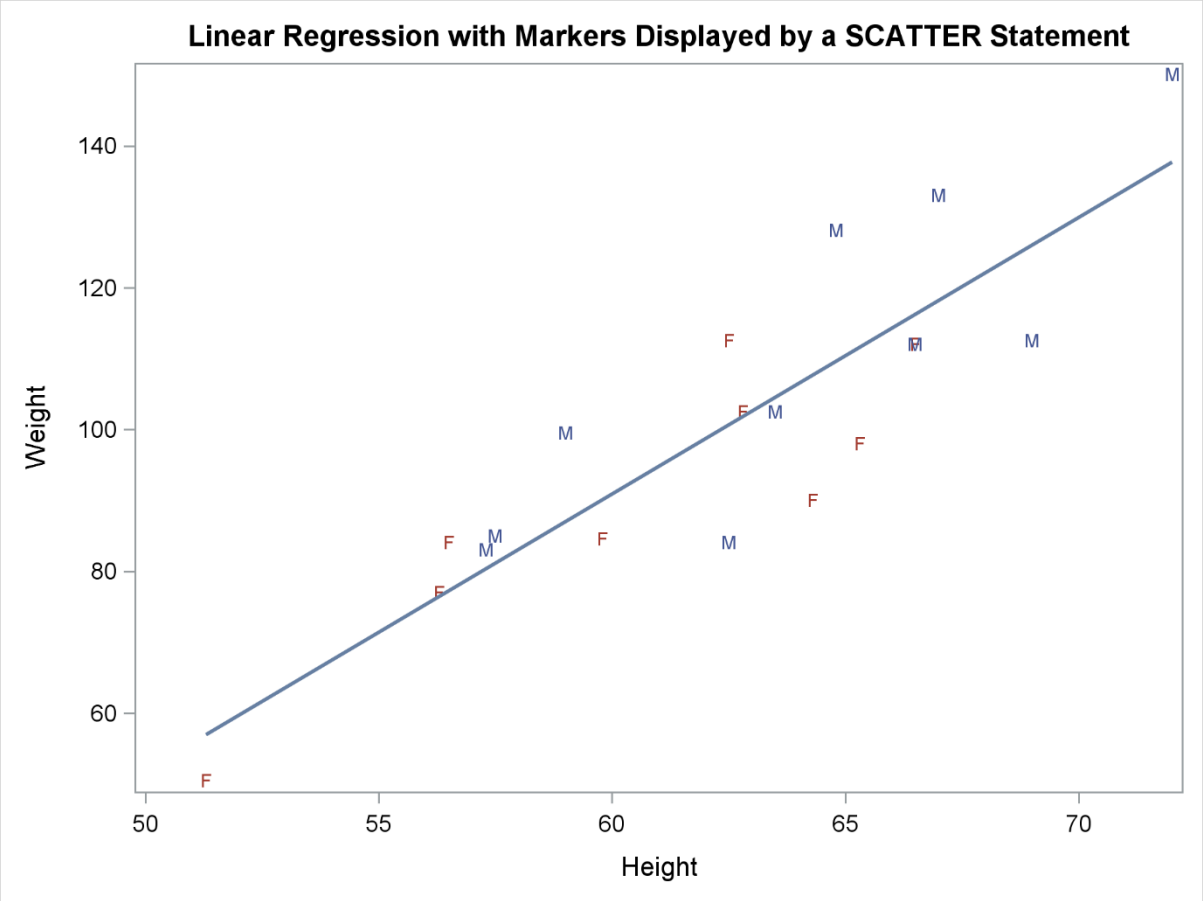
13. Is mean imputation of missing data acceptable practice?

Outliers data points will have a significant impact on the mean and hence, in such cases, it is not recommended to use the mean for replacing the missing values. Using mean values for replacing missing values may not create a great model and hence gets ruled out.

14. What is linear regression in statistics?

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.

Regression lines can be used as a way of visually depicting the relationship between the independent (x) and dependent (y) variables in the graph. A straight line depicts a linear trend in the data (i.e., the equation describing the line is of first order. For example, $y = 3x + 4$.



It will display the relationship between the dependent and independent variable that is there relationship so we go further to take that field to model building or not.

15. What are the various branches of statistics ?

Statistics is a study of presentation, analysis, collection, interpretation and organization of data

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 its not easy or possible to examine each member of the population.

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