1. Ans- a) True

2. Ans: a) Central Limit Theorem

3. Ans: - b) Modeling bounded count data

4. Ans: - d) All of the mentioned

5. Ans: - c) Poisson

6. Ans: - a) True

7.Ans: - b) Hypothesis

8. Ans: - a) 0 and c)1

9. Ans: - c) Outliers cannot confirm to the regression relationship

Q10

Ans: - The Normal Distribution, also called the Gaussian Distribution, is the most significant continuous probability distribution. We can also call this distribution as Bell Curve. A large number of random variables are either nearly or exactly represented by the normal distribution

Q11.

Ans: - Handling missing data is an important part of the data munging process that is integral to all data science projects. Incomplete observations can adversely affect the operation of machine learning algorithms so we must have procedures in place to properly manage this situation. Data imputation is one such procedure – it is the process of filling in missing values based on other data.

Common imputation methods of dealing with unknown and missing values are as follows:

1. Deleting Rows: -

This method commonly used to handle the null values. Here, we either delete a particular row if it has a null value for a particular feature and a particular column if it has more than 70-75% of missing values. This method is can only be used when there are enough samples in the data set. One has to make sure that after we have deleted the data, there is no addition of bias. Removing the data will lead to loss of information which will not give the expected results while predicting the output.

```
Data_name.dropna(inplace= True)

Data_name.isnull(). sum() #to check missing value
```

2. Replacing With Mean/Median/Mode

This strategy can be applied on a feature which has numeric data. We can calculate the mean, median or mode of the feature and replace it with the missing values. This is an approximation which can add variance to the data set. But the loss of the data can be negated by this method which yields better results compared to removal of rows and columns. Replacing with the above three approximations are a statistical approach of handling the missing values. This method is also called as leaking the data while training. Another way is to approximate it with the deviation of neighboring values. This works better if the data is linear.

```
data_name['Age'].isnull().sum() #finding null values in that column
data_name['Age'].mean()  #finding mean
    data_name['Age']. replace(np.NaN, data_name['Age'].mean()) # to replace
```

3. Assigning Unique Category:

A categorical feature will have a definite number of possibilities, such as gender, for example. Since they have a definite number of classes, we can assign another class for the missing values. Here, the features Cabin and Embarked have missing values which can be replaced with a new category, say, U for 'unknown'. This strategy will add more information into the dataset which will result in the change of variance. Since they are categorical, we need to find one hot encoding to convert it to a numeric form for the algorithm to understand it.

```
data_name['Name'].fillna('U'). head(2) # to feel unknown
```

4. Predicting the missing values

Using the features which do not have missing values, we can predict the nulls with the help of a machine learning algorithm. This method may result in better accuracy, unless a missing value is expected to have a very high variance. We will be using LinearRegression model to replace the null.

5. Using Algorithms Which Support Missing Values

The final technique is to do nothing. The majority of machine learning algorithms do not work with missing data.

Q12

Ans: - A/B testing is a basic randomized control experiment. It is a way to compare the two versions of a variable to find out which performs better in a controlled environment. Let's say you own a company and want to increase the sales of your product. Here, either you can use random experiments, or you can apply scientific and statistical methods. A/B testing is one of the most prominent and widely used statistical tools. In the above scenario, you may divide the products into two parts – A and B. Here A will remain unchanged while you make significant changes in B's packaging. Now, on the basis of the response from customer groups who used A and B respectively, we try to decide which is performing better.

It is a hypothetical testing methodology for making decisions that estimate population parameters based on sample statistics. The **population** refers to all the customers buying your product, while the **sample** refers to the number of customers that participated in the test.

In Hypothesis testing we have to make two hypotheses

Null Hypothesis (Original)

Alternative Hypothesis (Predicted)

example, the Alternative Hypothesis (Predicted) is- "**the conversion rate of newsletter B is higher than those who receive newsletter A"**.

"We collect evidence and Reject Null Hypothesis Testing"

Q13.

Answer: -Mean imputation of missing data is an acceptable practice only when the missing value proportion is not large enough. But when the missing values are large enough and you impute them with the mean, the standard errors will be lesser than what they actually would have been. Small standard errors can lead to small p-values and this can create problems for us, because some variables will start appearing significant, which are ideally not significant.

Q14.

Ans: Regression analysis helps in determining the cause-and-effect relationship between variables. It is possible to predict the value of other variables (called dependent variable) if the values of independent variables can be predicted using a graphical method or the algebraic method.

Graphical Method: involves drawing a scatter diagram with independent variable on X-axis and dependent variable on Y-axis. After that a line is drawn in such a manner that it passes through most of the distribution, with remaining points distributed almost evenly on either side of the line.

A regression line is known as the line of best fit that summarizes the general movement of data. It shows the best mean values of one variable corresponding to mean values of the other. The regression line is based on the criteria that it is a straight line that minimizes the sum of squared deviations between the predicted and observed values of the dependent variable.

Algebraic Method: Algebraic method develops two regression equations of X on Y, and Y on X.

Regression equation: ------> Y(dependent)=a+bX

15.

Ans: - 1. Population: A population is the complete set group of individuals, whether that group comprises a nation or a group of people with a common characteristic.

2. Sample: Subset of population.

3. Mean: More often termed as "average", the meaning is the number obtained by computing the sum of all observed values divided by the total number

4. Meadian: Median is the middle value when the given data are ordered from smallest to largest

5. Mode: The mode is the most frequent number present in the given data. There can be more than one mode or none depending on the occurrence of numbers

6.Variance: Variance is the averaged square difference from the Mean

7. Standard Deviation: Standard Deviation measures how spread out the numerical values are. It is the square root of variance.

8.Range: Difference between the highest and lowest observations within the given data points.

9. Inter Quartile Range:

- **Q1:** middle value in the first half of the ordered data points
- **Q2:** median of the data points
- **Q3:** middle value in the second half of the ordered data points
- **IQR:** given by Q3-Q1

10. Skewness

It gives us a measure of distortion from symmetry (skew). Depending on whether the left or right tail is skewed for given data distribution, skewness is classified into Positive and Negative skewness as illustrated below

11. Inferential Statistics

It involves mathematical estimates that allow us to infer on a pattern or trend based on the sample data sets of a larger population. Helps to generalize, conclude and predict a bigger population

12. Descriptive Statistics

It helps in understanding the basic features of the data by summarizing them in a numerical or graphical way. Facts regarding the data involved can be presented by descriptive analysis, however, any kind of generalization or conclusion is not possible.

>. Normal Distribution:

>. Central Limit Theorem

>. Hypothesis Testing

>. P-value