1. True

2. Central limit theorem

3. Modeling bounded count data

4. All of the mentioned

5. Poison Distribution

6. True

7.Hypothesis

8. 0

9. Outliners cannot conform to the regression relationship

10. Normal distribution, also known as the Gaussian 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. In graph form, normal distribution will appear as a bell curve.

The Normal distribution is the most common type of distribution assumed in technical stock market analysis and in other types of statistical analyses. The standard normal distribution has two parameters: the mean and the standard deviation. For a normal distribution, 68% of the observation are within +/- one standard deviation of the mean, 95% are within +/- two standard deviation, and 99.7% are within +- three standard deviation

The normal distribution model is motivated by the central limit theorem. this theory states that averages calculated from independent, identically distributed random variables have approximately normal distributions, regardless of the type of distribution from which the variables are sampled.

11. Handling missing data is done by imputation technique. Handling missing data is one of important things for data so it should be done in a way that has a best possible value for the place we are filling it.

Missed data is shown by data.isnull().sum()- we can get how much data is missing is columns.

There is some advanced technique which is used for imputation of missing data,

1. Iterative Imputer:

This method treats other columns which are not having null values it takes it as feature columns and which are having null value take it as label column and train on them.

Finally, it will predict the nan data & impute. It's just like a regression problem. Here, null column is label & non-null as features.

2. KNN imputer:

KNN imputer will try to find the relation with the other columns and impute the data according the relation with other columns.

3. Simple Imputer:

Simple imputer will take mean of a column and replace it with the missing value. Mean is default of strategy we can change it to mode or median as per our data classification.

Basic method is fillna, which is not good technique for a greater number of rows.

12. A/B testing (also known as bucket testing or split-run testing) is a user experience research methodology. A/B tests consist of a randomized experiment with two variants, A and B. It includes application of statistical hypothesis testing or "two-sample hypothesis testing" as used in the field of statistics. A/B testing is a way to compare two versions of a single variable, typically by testing a subject's response to variant A against variant B, and determining which of the two variants is more effective.

A/B testing is a shorthand for a simple randomized controlled experiment, in which two samples (A and B) of a single vector-variable are compared. These values are similar except for one variation which might affect a user's behaviour. A/B tests are widely

considered the simplest form of controlled experiment. However, by adding more variants to the test, its complexity grows.

Statistical tools :

1 ttest

2 ANOVA

3 chi-square test(chi2 test)

13. No, it's not good technique because of the following reasons:

1 Leads to an underestimate of standard deviation.

2 Distorts relationship between variables by "pulling" estimates of the correlation towards zero.

3. Mean imputation reduces the variance of the imputed variable.

4. It also shrink standard error, which invalidates most hypothesis test.

Alternative solution is KNN imputer and Iterative Imputer

14. Linear regression is a basic and commonly used type of predictive analysis. The overall idea of regression is to examine two things:

1 .Does a set of predictor variables do a good job in predicting an outcome (dependent) variable?

2. Which variables in particular are significant predictors of the outcome variable, and in what way do they–indicated by the magnitude and sign of the beta estimates–impact the outcome variable?

These regression estimates are used to explain the relationship between one dependent variable and one or more independent variables. The simplest form of the regression equation with one dependent and one independent variable is defined by the formula y = c + b*x

Where

Y = estimated dependent variable score

C = constant

B = regression coefficient

X=score on the independent variable.

We can use linear regression when these assumption are followed :

1. Linearity: The relationship between X and the mean of Y is linear.
2. Homoscedasticity: The variance of residual is the same for any value of X.
3. Independence: Observations are independent of each other.
4. Normality: For any fixed value of X, Y is normally distributed.

15 . Various branches of statistics are:

1. Descriptive statistics - Descriptive statistics are brief descriptive coefficients that summarize a given data set, which can be either a representation of the entire population or a sample of a population. Descriptive statistics are broken down into measures of central tendency and measures of variability (spread). Measures of central tendency include the mean, median, and mode, while measures of variability include standard deviation, variance, minimum and maximum variables, kurtosis and skewness.

2 Inferential Statistics – It helps in finding the conclusion regarding the population after analysis on the sample drawn in it.

3 Data Collection – Data collection is all about how the actual data is collected.