1. What is a hypothesis test?

Ans. Hypothesis testing is a statistical method used to determine if there is enough evidence in a sample data to draw conclusions about a population. It involves formulating two competing hypotheses, the null hypothesis (H0) and the alternative hypothesis (Ha), and then collecting data to assess the evidence.

1. How to find normality of data?

Ans>: There are various methods available to test the normality of the continuous data, out of them, most popular methods are Shapiro–Wilk test, Kolmogorov–Smirnov test, skewness, kurtosis, histogram, box plot, P–P Plot, Q–Q Plot, and mean with SD.

1. Explain Shapiro Wilk test, Lilliefors (Kolmogorov-Smirnov) test , Anderson Darling test

Ans.: The Shapiro-Wilk test is a hypothesis test that is applied to a sample with a null hypothesis that the sample has been generated from a normal distribution. If the p-value is low, we can reject such a null hypothesis and say that the sample has not been generated from a normal distribution.

* The Lilliefors test uses the same calculations as the Kolmogorov-Smirnov test, but the table of critical values in the Lilliefors Test Table is used instead of the Kolmogorov-Smirnov Table. Since the critical values in this table are smaller, the Lilliefors Test is less likely to show that data is normally distributed.
* The Anderson-Darling test (Stephens, 1974) is used to test if a sample of data came from a population with a specific distribution. It is a modification of the Kolmogorov-Smirnov (K-S) test and gives more weight to the tails than does the K-S test.

1. What are the different types of Hypothesis testing?

Ans.: There are three types of hypothesis tests: right-tailed, left-tailed, and two-tailed. When the null and alternative hypotheses are stated, it is observed that the null hypothesis is a neutral statement against which the alternative hypothesis is tested.

1. Explain one sample t- test. And its assumptions

Ans.:

The one sample t test, also referred to as a single sample t test, is a statistical hypothesis test used to determine whether the mean calculated from sample data collected from a single group is different from a designated value specified by the researcher.

The one sample t-test has four main assumptions: The dependent variable must be continuous (interval/ratio). The observations are independent of one another. The dependent variable should be approximately normally distributed.

1. Explain independent sample t- test. And its assumptions

Ans: The independent samples t-test is used to compare two sample means from unrelated groups. This means that there are different people providing scores for each group. The purpose of this test is to determine if the samples are different from each other.

Independent samples t -test Assumptions:

The data are numeric. Observations are independent of one another (that is, the sample is a simple random sample and each individual within the population has an equal chance of being selected) The sample mean, ¯¯¯¯¯X , is normally distributed. Equal variances between groups

1. Explain paired t- test. And its assumptions

Ans: A paired t-test (also known as a dependent or correlated t-test) is a statistical test that compares the averages/means and standard deviations of two related groups to determine if there is a significant difference between the two groups.

Paired t-test assumptions:

Subjects must be independent. Measurements for one subject do not affect measurements for any other subject. Each of the paired measurements must be obtained from the same subject. For example, the before-and-after weight for a smoker in the example above must be from the same person.

1. What is level of significance

Ans:The level of significance is the measurement of the statistical significance. It defines whether the null hypothesis is assumed to be accepted or rejected. It is expected to identify if the result is statistically significant for the null hypothesis to be false or rejected.

1. What is confidance interval

Ans.: A confidence interval displays the probability that a parameter will fall between a pair of values around the mean. Confidence intervals measure the degree of uncertainty or certainty in a sampling method.

1. What is the difference between an error of type I and an error of type II?

Ans.: a Type I error means rejecting the null hypothesis when it's actually true, while a Type II error means failing to reject the null hypothesis when it's actually false.

1. What is p value?

Ans.: The P value is defined as the probability under the assumption of no effect or no difference (null hypothesis), of obtaining a result equal to or more extreme than what was actually observed. The P stands for probability and measures how likely it is that any observed difference between groups is due to chance.

1. Explain probability sampling methods?

Ans.: Probability sampling means that every member of the population has a chance of being selected. It is mainly used in quantitative research. If you want to produce results that are representative of the whole population, probability sampling techniques are the most valid choice.

1. What do you mean by Time Series Analysis?

Ans.: Time series analysis is a specific way of analyzing a sequence of data points collected over an interval of time. In time series analysis, analysts record data points at consistent intervals over a set period of time rather than just recording the data points intermittently or randomly.

1. How do you define Normal Distribution?

Ans.: 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.

1. How do you treat outliers in a dataset?

Ans.: How to deal with outliers? Three main methods of dealing with outliers, apart from removing them from the dataset: 1) reducing the weights of outliers (trimming weight) 2) changing the values of outliers (Winsorisation, trimming, imputation) 3) using robust estimation techniques (M-estimation).

1. What Is Linear Regression?

Ans: Linear regression is a data analysis technique that predicts the value of unknown data by using another related and known data value. It mathematically models the unknown or dependent variable and the known or independent variable as a linear equation.

1. What Is Logistic Regression?

Ans.: A logistic regression model predicts a dependent data variable by analyzing the relationship between one or more existing independent variables. For example, logistic regression could be used to predict whether a political candidate will win or lose an election or whether a high school student will be admitted to a particular college. These binary outcomes enable straightforward decisions between two alternatives.

1. Explain ANOVA test.

Ans: ANOVA stands for Analysis of Variance. It is a statistical method used to analyze the differences between the means of two or more groups or treatments. It is often used to determine whether there are any statistically significant differences between the means of different groups.

1. Explain one way ANOVA , two way ANOVA Test

Ans.: One-way ANOVA is a statistical method to test the null hypothesis (H0) that three or more population means are equal vs. the alternative hypothesis (Ha) that at least one mean is different.

A two-way ANOVA is used to estimate how the mean of a quantitative variable changes according to the levels of two categorical variables. Use a two-way ANOVA when you want to know how two independent variables, in combination, affect a dependent variable.

1. What is Wilcoxon Test

Ans.: he Wilcoxon test, which can refer to either the rank sum test or the signed rank test version, is a nonparametric statistical test that compares two paired groups. The tests essentially calculate the difference between sets of pairs and analyze these differences to establish if they are statistically significantly different from one another.

1. What is Mann Whitney test

Ans.: The Mann-Whitney U test, also known as the Mann-Whitney-Wilcoxon test, is a non-parametric statistical test used to compare two independent groups when the dependent variable is ordinal or continuous, but not normally distributed. It's often used as an alternative to the independent samples t-test when the assumptions of normality and equal variances are not met.

1. What is Kruskal Wallis Test

Ans.: It is a nonparametric method for testing whether samples are originated from the same distribution. It extends the Mann-Whitney U test to more than two groups. The null hypothesis of the Kruskal-Wallis test is that the mean ranks of the groups are the same.

1. Explain chi square test

Ans.: A chi-square test is a statistical test that is used to compare observed and expected results. The goal of this test is to identify whether a disparity between actual and predicted data is due to chance or to a link between the variables under consideration.

1. Explain Regression analysis and its applications.

Ans.: Regression analysis is helpful in financial forecasting to model relationships between financial variables, such as stock prices and economic indicators. It can help identify trends, estimate future values, and manage financial risk by analyzing historical data and making informed predictions based on relevant factors.

1. Explain linear regression, multiple regression , stepwise regression.

Ans.: Linear regression attempts to establish the relationship between the two variables along a straight line. Multiple regression is a type of regression where the dependent variable shows a linear relationship with two or more independent variables.

Stepwise regression is a method that iteratively examines the statistical significance of each independent variable in a linear regression model. The forward selection approach starts with nothing and adds each new variable incrementally, testing for statistical significance.