1) Differentiate between Inferential Statistics and Descriptive Statistics.

Descriptive Statistics:

1. It is a summary statistic that quantitatively describes or summarizes the feature of a collection of information. It helps in knowing the data better.
2. It gives information about raw data which describes the data in some manner.
3. It helps in organizing, analyzing, and to present data in a meaningful manner.
4. It is used to describe a situation.
5. It explains already known data and is limited to a sample or population having a small size.
6. It can be achieved with the help of charts, graphs, tables, etc.

Inferential Statistics

1. It is defined as inferring population with the help of sample(true representation or a subset of population)
2. It makes inferences about the population using data drawn from the population.
3. It allows us to compare data, and make hypotheses and predictions.
4. It is used to explain the chance of occurrence of an event.
5. It attempts to reach the conclusion about the population.
6. It can be achieved by probability.

2) Differentiate between population and sample.

A **population** is the entire group that you want to draw conclusions about. Universe of elements to be studied. It can be classified according to the number of individuals that make it up. It has statistical variables. To analyze the data collected regarding the common characteristics shared by the elements for various purposes.

A **sample** is the specific group that you will collect data from. The size of the sample is always less than the total size of the population. Selection of a part of the population. It is part of the population: it should comprise between 5% and 10% to be most effective. Variable could be random. To study the behavior, characteristics, tastes, or properties of a representative part of the population.

3) What is a hypothesis?Differentiate between null and alternative hypothesis.

A **hypothesis** is a testable prediction which is expected to occur. It can be true or false based on the underlying information in the data provided for the testing.

1. **Alternate Hypothesis:**
   In the Alternate hypothesis, there is no relationship between the two variables. Generally, researchers and scientists try to reject or disprove the null hypothesis. If the null hypothesis is accepted researchers have to make changes in their opinions and statements. This hypothesis is denoted by H1 or Ha. It is generally used when we reject the null hypothesis. In this hypothesis, the p-value is smaller than the significance level.
2. **Null Hypothesis:** In the null hypothesis, there is some relationship between the two variables i.e. They are dependent upon each other. Generally, researchers and scientists try to accept or approve the null hypothesis. If the alternative hypothesis gets accepted researchers do not have to make changes in their opinions and statements. This hypothesis is denoted by H0. It gets accepted if we fail to reject the null hypothesis. In this hypothesis, the p-value is greater than the significance level.

4) What is the central limit theorem?

Central limit theorem is a statistical theory which states that when the large sample size has a finite variance, the samples will be normally distributed and the mean of samples will be approximately equal to the mean of the whole population.

5) Differentiate between type I and type II errors.

1. **Type I error (false-positive)**
   occurs if an investigator rejects a null hypothesis that is actually true in the population
   - It is also known as a false-positive.
   - It occurs if the researcher rejects a correct null hypothesis in the population i.e., incorrect rejection of the null hypothesis.
   - Measured by alpha (significance level).
   - If the significance level is fixed at 5%,it means there are about five chances of type – 1 error out of 100.
2. **Type II error (false-negative)**
   occurs if the investigator fails to reject a null hypothesis that is actually false in the population.
   - It is also known as a false negative.
   - It occurs if a researcher fails to reject a null hypothesis that is actually a false hypothesis.
   - Measured by beta (the power of test).
   - The probability of committing a type -2 error is calculated by 1 – beta (the power of test).

6) What is linear regression?

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. For instance, suppose that you have data about your expenses and income for last year. Linear regression techniques analyze this data and determine that your expenses are half your income. They then calculate an unknown future expense by halving a future known income.

7) What are the assumptions required for linear regression?

1. **Linear relationship:** There exists a linear relationship between the independent variable, x, and the dependent variable, y.
2. **Independence:**
   The residuals are independent. In particular, there is no correlation between consecutive residuals in time series data.
3. **Homoscedasticity:**
   The residuals have constant variance at every level of x.
4. **Normality:**
   The residuals of the model are normally distributed.

8) How is the statistical significance of an insight assessed?

Hypothesis testing is guided by statistical analysis. Statistical significance is calculated using a p-value, which tells you the probability of your result being observed, given that a certain statement (the null hypothesis) is true. If this p-value is less than the significance level set (usually 0.05), the experimenter can assume that the null hypothesis is false and accept the alternative hypothesis. Using a simple t-test, you can calculate a p-value and determine significance between two different groups of a dataset.

9) What Is Mean?

Mean is the average of the numerical observations which is equal to the sum of the observations divided by the number of observations. Thus, the mean is a number around which the entire data set is spread.

10) What is the meaning of standard deviation?

Standard deviation measures the spread of a data distribution. It measures the typical distance between each data point and the mean. It depicts the closeness of the data point to the mean and is calculated as the square root of the variance. In data science, the standard deviation is usually used to identify the outliers in a data set. The data points which lie one standard deviation away from the mean are considered to be unusual.

11) What is correlation?

Correlation: Correlation refers to a process for establishing the relationships between two variables. Correlation methods summarise the relationship between two variables in a single number called the correlation coefficient. The correlation coefficient is usually represented using the r, ranging from -1 to +1.

A correlation coefficient close to plus 1 means a **positive relationship** between the two variables, with increases in one of the variables being associated with increases in the other variable. Positive Correlation – Same direction.

A correlation coefficient close to -1 indicates a **negative relationship** between two variables, with an increase in one of the variables being associated with a decrease in the other variable. Negative Correlation – Opposite direction

Scatter Diagram:A scatter diagram is a diagram that shows the values of two variables, X and Y, along with the way in which these two variables relate to each other. The values of variable X are given along the horizontal axis, with the values of variable Y given on the vertical axis.

12) What is the meaning of covariance?

Covariance is a measure of the relationship between two random variables and to what extent, they change together. Or we can say, in other words, it defines the changes between the two variables, such that change in one variable is equal to change in another variable.

1. **Positive Covariance**
   If the covariance for any two variables is positive, that means, both the variables move in the same direction. Here, the variables show similar behaviour. That means, if the values (greater or lesser) of one variable corresponds to the values of another variable, then they are said to be in positive covariance.
2. **Negative Covariance**
   If the covariance for any two variables is negative, that means, both the variables move in the opposite direction. It is the opposite case of positive covariance, where greater values of one variable correspond to lesser values of another variable and vice-versa.

13) Where is inferential statistics used?

1. Making estimates about populations (for example, the mean SAT score of all 11th graders in the US).
2. Testing hypotheses to draw conclusions about populations (for example, the relationship between SAT scores and family income).

14) What is one sample t-test?

The **one-sample t test** is a statistical procedure used to compare a mean value measured in a sample to a known value in the population. It is specifically used to test hypotheses concerning the mean in a single population with an unknown variance.

15) What is the relationship between standard deviation and standard variance?

Variance is nothing but the average taken out of the squared deviations.

Standard Deviationis defined as the root of the mean square deviation.

16) What is a one way ANOVA test?

One-Way ANOVA ("analysis of variance") compares the means of two or more independent groups in order to determine whether there is statistical evidence that the associated population means are significantly different. One-Way ANOVA is a parametric test. This test is also known as: One-Factor ANOVA