1. **Difference Between Descriptive statistics and Inferential Statistics**

**Descriptive statistics:**

Organizes, describes and presents data in a meaningful way with the help of charts and graphs.

It explains already known data and limited to sample.

[Measures of central tendency](https://www.cuemath.com/data/measures-of-central-tendency/): mean, median and mode

Measures of variability : standard deviation and range

This is used to describe a situation

Less error

**Inferential Statistics:**

Tests, predicts, and compares data obtained from various samples.

It conclude about population.

Hypothesis testing and regression analysis

This is used to explain the chances of occurrence of an event.

More error

**2**. **Difference between population and sample.**

**Population:**

Population includes all the elements from the data set.

Identifying the characteristics.

Ex. All employees in an office

**Sample:**

A small group taken from population.

Making conclusion about population.

Ex. Out of all the employees, all managers in the office.

1. **What is hypothesis? Differentiate between null and alternative hypothesis**.

Hypothesis is an assumption that may be true or false. Hypothesis should be based on existing theories and knowledge. It is easy to understand and easy to test. It also tells about relation between variables.

Smoking leads to cancer is example of hypothesis.

**Null hypothesis:**

It suggests that there is no relationship between the two variables.

It is generally used when we reject the null hypothesis

Here the p-value is smaller than the significance level.

This hypothesis is denoted by H0.

**Alternative hypothesis:**

 Here is some relationship between the two variables.

It gets accepted if we fail to reject the null hypothesis.

Here the p-value is greater than the significance level.

This hypothesis is denoted by Ha or H1.

**4. What is central limited theorem?**

Relationship between shape of population distribution and shape of sampling distribution of mean is known as central limited theorem.

It states that the sampling distribution of the sample means approaches a normal distribution as the sample size gets larger.

It holds True when sample size>30 .

1. **Difference between type I and Type II error.**

**Type I error:**

Type I error is a false positive conclusion.

It means rejecting the null hypothesis when it’s actually true.

Null hypothesis is that when there Is not any relation btwn variables

Ex. If you get tested for covid and test result says you have coronavirus, but you actually don’t.

**Type II error:**

Type II error is a false negative conclusion.

It  means not rejecting the null hypothesis when it’s actually false.

Ex. If you get tested for covid and test result says you don’t have coronavirus, but you actually do.

1. **What is linear regression?**

A statistical way of measuring the relationship between variables.

It analyze the specific relationship between the independent and dependent variable.

Ex. Relation between qualification and income.

There are two types of linear regression- Simple and Multiple.

**Simple Linear Regression:**

If a single independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Simple Linear Regression.

**Multiple Linear Regression:**

If more than one independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Multiple Linear Regression.

1. **What are the assumptions required for linear regression?**

**Linear relationship between the features and target:**  
Linear regression assumes the linear relationship between the dependent and independent variables.

**Small or no multicollinearity between the features:**  
Multicollinearity means high-correlation between the independent variables. Due to multicollinearity

**Homoscedasticity Assumption:**  
Homoscedasticity is a situation when the error term is the same for all the values of independent variables.

**Normal distribution of error terms:**  
Linear regression assumes that the error term should follow the normal distribution pattern.

**No autocorrelations:  
The linear regression model assumes no autocorrelation in error terms.**

1. **How is the statistical significance of an insight assessed?**

Statistical significance can be accessed using hypothesis testing:

– Stating a null hypothesis which is usually the opposite of what we wish to test (classifiers A and B perform equivalently, Treatment A is equal of treatment B)  
 Then, we choose a suitable statistical test and statistics used to reject the null hypothesis.

Also, we choose a critical region for the statistics to lie in that is extreme enough for the null hypothesis to be rejected (p-value)

We calculate the observed test statistics from the data and check whether it lies in the critical region  
Common tests:  
– One sample Z test  
– Two-sample Z test  
– One sample t-test  
– paired t-test  
– **Two** sample pooled equal variances t-test  
– Two sample unpooled unequal variances t-test and unequal sample sizes (Welch’s t-test)  
– Chi-squared test for variances  
– Chi-squared test for goodness of fit  
– Anova (for instance: are the two regression models equals? F-test)  
– Regression F-test (i.e: is at least one of the predictor useful in predicting the response?)

**9.What is mean?**

In statistics, the mean is one of the measures of central tendency, apart from the mode and median. Mean is nothing but the average of the given set of values. It denotes the equal distribution of values for a given data set.

**10.what is the meaning of Standard Deviation?**

**Standard Deviation** is a measure which shows how much variation (such as spread, dispersion, spread,) from the mean exists. The standard deviation indicates a “typical” deviation from the mean. It is a popular measure of variability because it returns to the original units of measure of the data set. *It is* represented by the symbol ‘σ.’

**11.what is correlation?**

Correlation is a statistical measure that indicates the extent to which two or more variables fluctuate in relation to each other. A [positive correlation](https://www.techtarget.com/whatis/definition/positive-correlation) indicates the extent to which those variables increase or decrease in parallel; a [negative correlation](https://www.techtarget.com/whatis/definition/negative-correlation) indicates the extent to which one variable increases as the other decreases. A [correlation coefficient](https://www.techtarget.com/whatis/definition/correlation-coefficient) is a statistical measure, of the degree to which changes to the value of one variable predict change to the value of another.

**12.what is the meaning of covariance?**

Covariance is a statistical tool used to determine the relationship between the movements of two random variables.

When two stocks tend to move together, they are seen as having a positive covariance; when they move inversely, the covariance is negative.

**13.Where is inferential statistics used?**

Inferential statistics is used in a wide range of fields to make conclusions or predictions about a population based on a sample of data. Here are some common examples:

Medical research: Inferential statistics is used to make predictions about the effectiveness of a new drug or treatment on a population, based on the results of a clinical trial on a sample of patients.

Market research: Inferential statistics is used to make predictions about consumer behavior, market trends, and product demand, based on data from a sample of customers.

Social sciences: Inferential statistics is used in fields such as psychology, sociology, and education to study human behavior and make predictions about population-level trends and patterns.

Sports: Inferential statistics is used to make predictions about the outcome of games and tournaments, and to analyze player performance and team dynamics.

**14. What is one sample t-test?**

The one-sample t-test is a statistical hypothesis test used to determine whether an unknown population mean is different from a specific value.

You can use the test for continuous data. Your data should be a random sample from a normal population.

The One Sample *t* Test can only compare a single sample mean to a specified constant. It can not compare sample means between two or more groups. If you wish to compare the means of multiple groups to each other, you will likely want to run an Independent Samples *t* Test (to compare the means of two groups) or a One-Way ANOVA (to compare the means of two or more groups).

**15.What is the relationship between standard deviation and standard variance?**

The standard deviation is the square root of the variance.

Variance is a measure of the spread of a set of numbers around the mean, while standard deviation is the square root of variance and is expressed in the same units as the original data. So standard deviation gives a more interpretable value than variance as it's in the same units as the original data, while variance is squared units.

**16.What is a one way anova test?**

**ANOVA**, which stands for Analysis of Variance, is a [statistical test](https://www.scribbr.com/statistics/statistical-tests/) used to analyze the difference between the [means](https://www.scribbr.com/statistics/mean/) of more than two groups.

One-way ANOVA (Analysis of Variance) is a statistical hypothesis test that is used to compare the means of three or more independent groups/samples. The goal of this test is to determine if there is any significant difference in the mean of the groups. It assumes that the samples are independent and have approximately equal variances.

In a one-way ANOVA test, the null hypothesis is that the means of all groups are equal, and the alternative hypothesis is that at least one group mean is different from the others. The test statistic used in the one-way ANOVA test is the F-statistic, which is calculated based on the ratio of the explained variance (the variance between the groups) and the unexplained variance (the variance within the groups). If the calculated F-statistic is greater than the critical value from the F-distribution, the null hypothesis is rejected, and it is concluded that there is evidence of a difference in means among the groups.