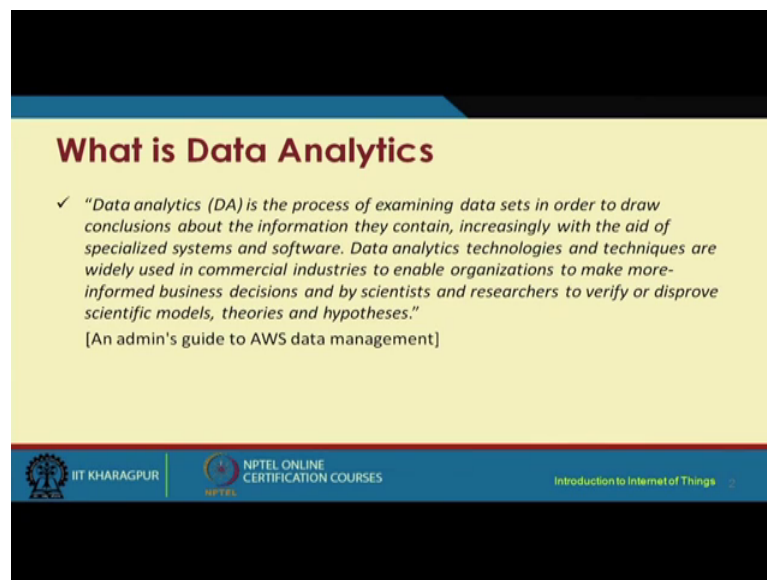**Introduction to Internet of Things**
Prof. Sudip Misra
Department of Computer Science & Engineering
Indian Institute of Technology, Kharagpur

Lecture - 56
Data Handling and Analytics- Part- II

In this lecture on data analysis and sorry Data Handling and Analytics. In the first part we will focused mostly on data handling, and in the second part we are going to focus mostly on the analytics.

So, having captured the data and storing it in the cloud or in the server or whatever storage mechanisms we have. We now have to use the data, for using it we have to analyze it: we have to analyze the data. So, from this there are different tools, different methodologies that they are; the most common the most primitive once are based on statistical methods; so basic statistical methods can be applied on the store data in what order to make more sense out of that data in order to get more insight into that data it is stored.

(Refer Slide Time: 01:20)

So, we have data analytics. Basically, the data have to be analyzed. So, in the context of IoT people talk a lot about data analytics. So, what is the state analytics? I am going to read one of these definitions. So, data analytics is the process of examining the data sets in order to draw conclusions about the information they contain. That means, what

information is contained in this data sets. Increasingly with the aid of specialized systems and software; so with different specialized software, systems, etcetera to get insight into the data that is existing.

Data analytics technologies and techniques are widely used in commercial industries to enable organizations to make more informed business decisions and by scientist and researchers to verify or disprove the scientific models theories and hypothesis. So, this is basically the premise in which data analytics basically work and the different concerns of data analytics are basically: I mentioned in this particular definition.
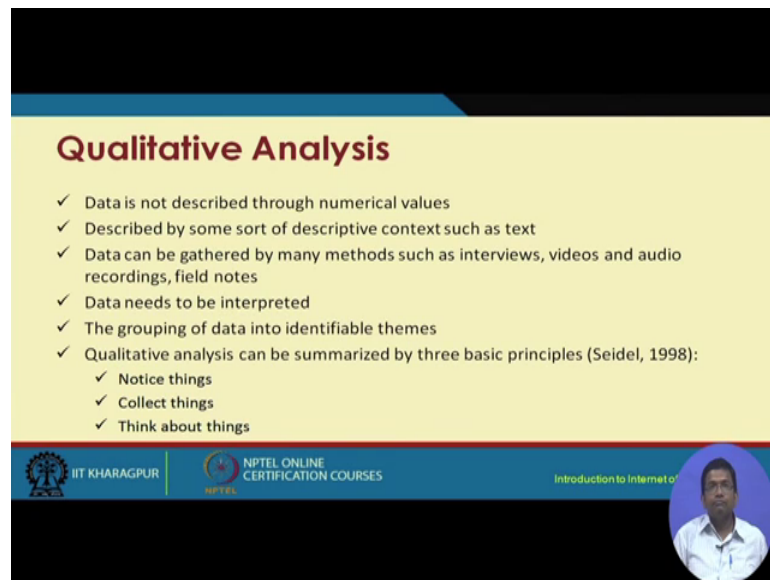
(Refer Slide Time: 02:39)

So, when we talk about analysis in general broadly analysis comes in two forms. So, we can either perform qualitative analysis on the data that has been obtained or we can perform quantitative analysis. So, qualitative analysis basically deals with the analysis of data that are categorical in nature- so qualitative analysis. Whereas, quantitative analysis refers to the process by which numerical methods can be used; numerical data can be analyzed through quantitative analysis.

So, categorical data: qualitative analysis is good enough, for numerical data quantitative analysis quantitative methods are used for.

(Refer Slide Time: 03:31)

So, qualitative analysis data is not described through numerical values, but are described by some sort of descriptive contexts such as text. This qualitative data can be gathered by many methods such as from interviews, interviewing different people, from videos, from audio recordings, field notes you know industry manuals and so on.

This data needs to be interpreted; the grouping of the data can be where should be performed in to identifiable themes in quantitative qualitative analysis. And the qualitative analysis can be summarized by three basic principles. Notice the things collect the things and think about it. We do not need to get into details of each of these.

Next one is the quantitative analysis. So, quantitative analysis is on the numeric data using different statistical methods such as descriptive statistics, more specifically finding out the mean of the dataset median standard deviation and so on. The following are often involved with the quantitative analysis, statistical models and analysis of variance, then data dispersion analysis of relationship between variables, contingency and correlation, then regression analysis, statistical significance, precision, error limits and so on.
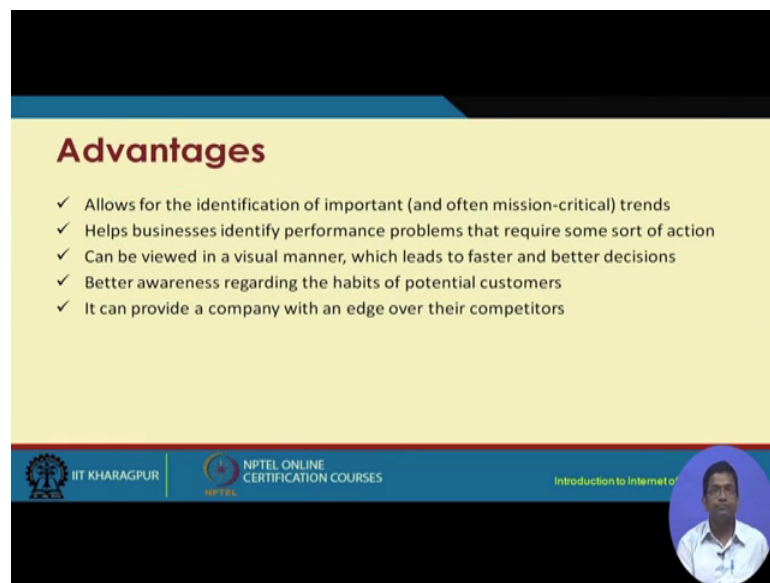
So, these are the different quantitative methods that are used for quantitative analysis of data.

Now comes the comparison between qualitative data and quantitative data. So, qualitative data can be mostly observed, whereas quantitative data can be measured. Qualitative data involves descriptions its more qualitative it involves descriptions, on the other hand quantitative data involves numbers, numeric's and so on. Whereas, in qualitative data the emphasis is on quality, in the quantitative data the emphasis is on quantity. Examples of qualitative data include colour, smell, taste, etcetera which cannot be quantified so easily. On the other hand quantifiable data include volume, weight, etcetera; these numbers these are figures which can be used to perform different numerics.

(Refer Slide Time: 06:14)

The advantages of data analytics is that: it allows for the identification of important trends, it helps the businesses identify performance problems that we require some sort of action- some prediction can be performed. By analysis of the data fast data something you know, so the businesses they can understand that what has gone wrong quantitatively we can be done or even qualitatively. So, data analytics are useful for that.

So, the analytics can also be performed in a visual manner and that can help in faster and better decision making. Analytics can provide a company with an edge over their competitors.
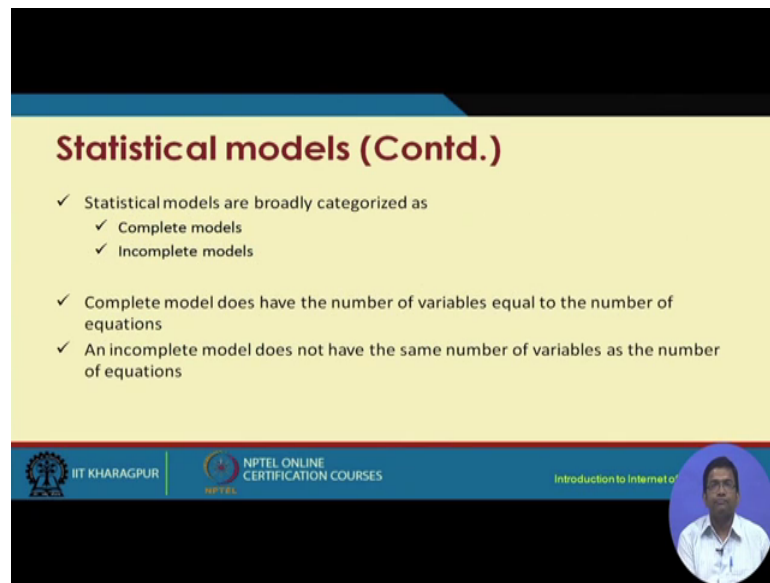
(Refer Slide Time: 07:00)

So, that is the reason actually data analytics has become very popular in the industry, not only in the industrial almost in all spheres of life data analytics has become very popular. And because you can get more insight into what is going on in the processes that are occurring around you.

Statistical different models of statistical; statistical models can be adopted in order to perform quantitative analysis. And a statistical model can is defined as the mathematical equation that is formulated to form the relationship between variables. A statistical model illustrates how a set of random variables is related to another set of random variables. And it is a statistical model is represented as an ordered pair X P; where X denotes the set of all possible observations and P refers to the set of probability of distributions on X.
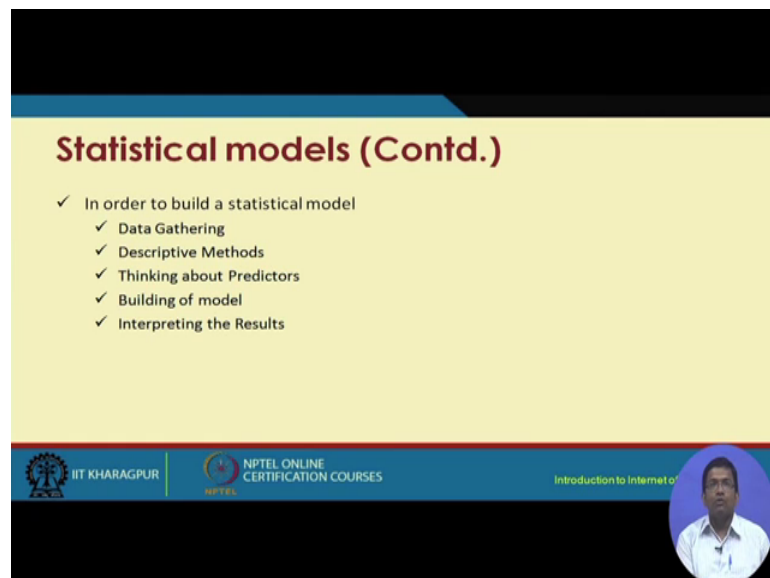
(Refer Slide Time: 08:05)

Statistical models are broadly categorized as complete models and incomplete models. Complete models have the same number of variables as the number of equations. So, the number of equations and the number of variables in the complete models are the same. So, if we have the number of variables equating with the number of equations what we have is a complete model. And in an incomplete model the number of variables and the number of equations are not the same- they do not match.

(Refer Slide Time: 08:39)

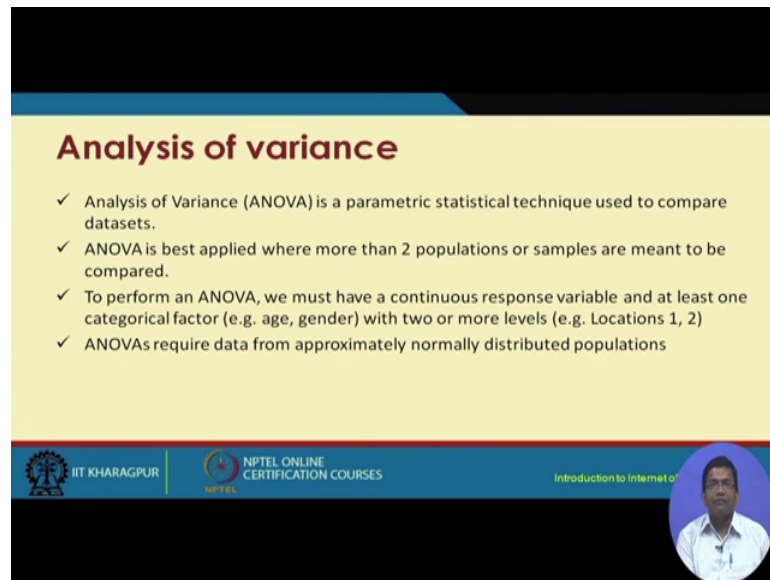So, in order to build a statistical model it is required to gather the data, perform descriptive methods, think about what are the predictors, then build the model and then interpret the results.
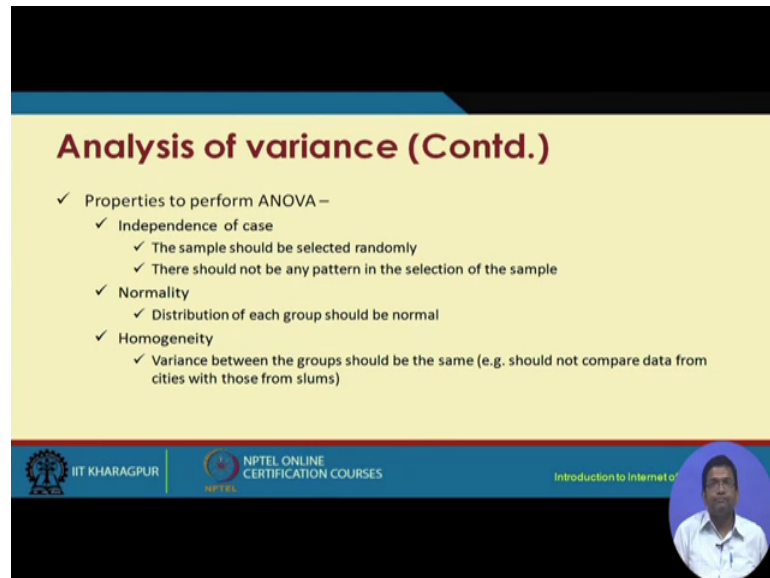
(Refer Slide Time: 08:55)

Analysis of variance in short known as ANOVA analysis: is a parametric statistical technique that can be used to compare two data sets- two or more data sets they can be compared. So, ANOVA is best applied when more than two populations of samples are meant to be compared. So, we have one dataset, we have another dataset, we want to compare these two populations to see that how much is the correlation between these two datasets, what sort of similarity exits between these two database sets.

So, to perform ANOVA one has to have a continuous response variable and at least one categorical factor. For example, age gender, etcetera, with at least two or more levels example location 1, location 2, etcetera. So, what it means is basically levels mean that one location: one location Kharagpur another location Kolkata. So, these are two different locations corresponding to two different levels. And categories mean the age is one category. So, with respect to a particular category like age you know. So, at two different locations what is the similarity or what is the dissimilarity; similarly with respect to gender or any other category.

ANOVA requires data from approximately normally distributed population. So, this is a very important assumption or a very important requirement that you know. So, normal

distribution is required for performing; normal distribution of the data set has to be there in order to perform ANOVA analysis.

(Refer Slide Time: 10:42)

The properties to perform ANOVA: one is the independence of case, the sample that is selected should be random; random is selected there should not be any bias, there should not be any pattern, in the selected sample. Normality is the second property which constant the distribution of each group should be normal, so normal distribution of the data within the group; and homogeneity which constants variance between the groups and the variance should be the same. So, we should not have a scenario to compare the data from cities with the data from maybe slums areas or maybe the data of Kharagpur compared with the data of Kolkata.

So, because we have a town we have a city so two different datasets you know compare of comparing with each other. So, they have huge variance. And the variance should be as much minimal as possible.

(Refer Slide Time: 11:50)

Analysis of variance has three different types: one way analysis which constant one fixed factor. For example, the factors could be age, gender, etcetera. Could be two way analysis where two or more or two factors are going to be involved. So, both maybe both age and gender will be considered in a two way ANOVA analysis. And it can be k way analysis where k factor variables are involved.
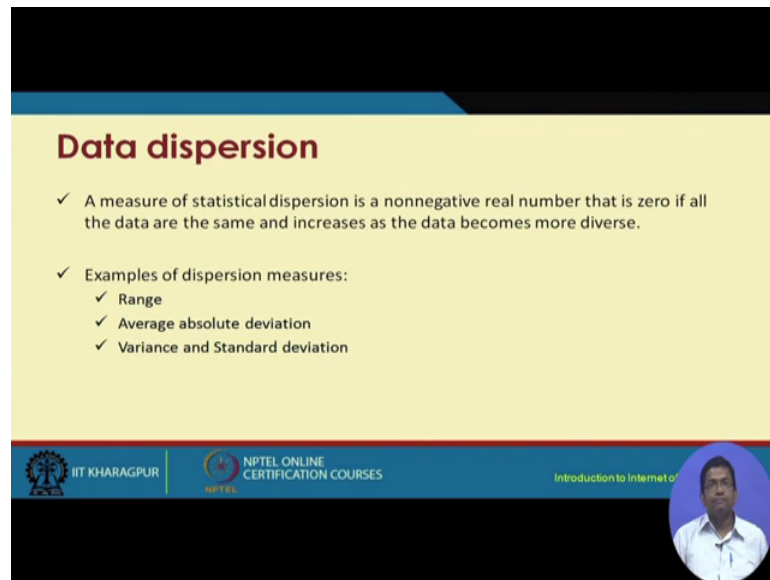
(Refer Slide Time: 12:21)

Then there are different ways, different features that are there for performing analysis of variance- total sum of square is 1, f ratio is another and the degree of freedom. So, all of

these things have to be taken into consideration in order to perform and its variance. I am not going to put through these, because this is not essentially a course in statistical methods. And these are all available, but what is important is that ANOVA analysis can be used in order to perform analytics on the data that is obtained from IoT systems.

(Refer Slide Time: 13:01)

The next concept that has to be understood is data dispersion. Data dispersion concerns how much is the dispersion; that means, dispersion is basically a measure of the statistical. So, it is a measure of statistical dispersion a non negative real number that is 0 if all the data are the same and it increases as the data becomes more diverse. Examples of dispersion measures include: range, average, absolute deviation, variance, and standard deviation. So, typically when we talk about dispersion we typically talked about in terms of variance and standard deviation.

(Refer Slide Time: 13:58)

So, how much the deviate from the norm? So, this is what dispersion constants. So, here as I already mentioned range and what is meant by it the absolute standard deviation is given. So, the average of the absolute deviation is given, variance and standard deviation here well known methods and (Refer Time: 14:16) dispersion deciding. And these are given over here.
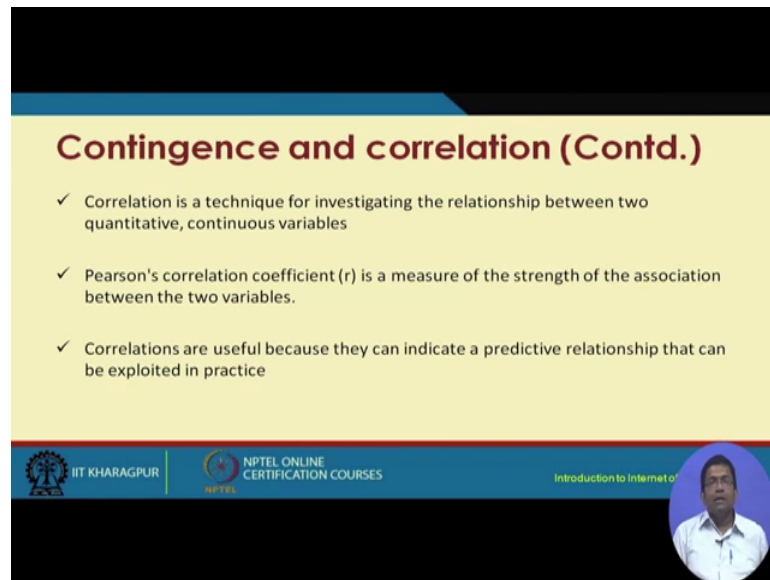
(Refer Slide Time: 14:19)

Next comes a contingence and correlation. So, in statistics a contingency table is a type of table in a matrix format that displays the multivariate frequency distribution of the

variables. It provides the basic picture of the interrelation between two variables. Correlation is a technique for investigating the relationship between two continuative continuous variables.
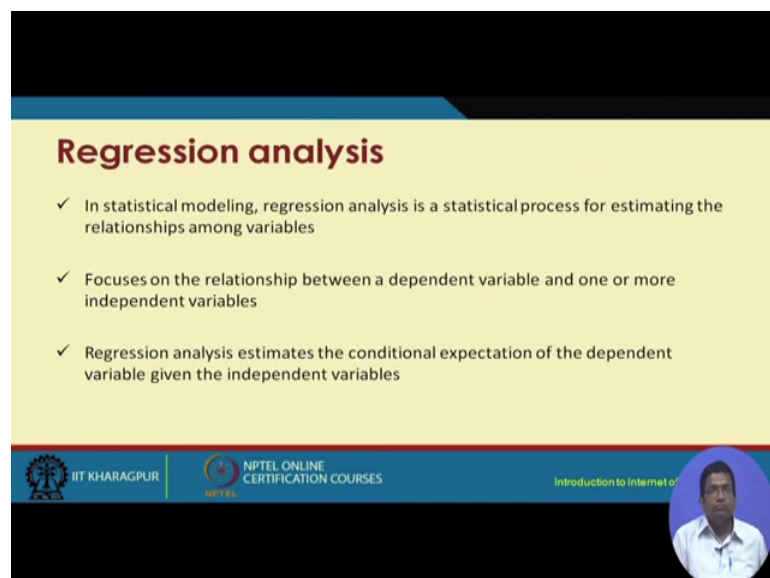
(Refer Slide Time: 14:42)

So, they have to be continuous variable this is very important. And how much they are correlated these two variables how much they are correlated and what is the relationship between them. So, a popular measure is the Pearson's correlation coefficient. And it basically measures the strength of association between two variables.
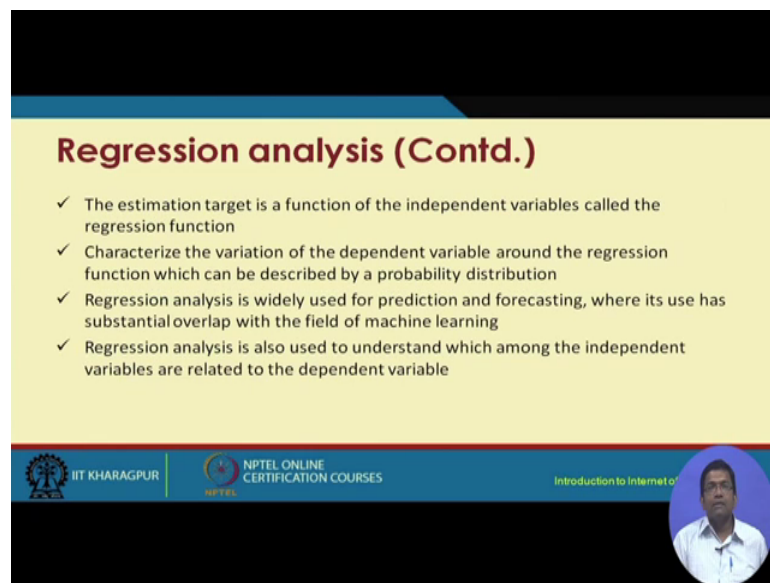
(Refer Slide Time: 15:13)

Then comes regression analysis. So, regression analysis basically tries to estimate the relationship among the different variables. It focuses on the relationship between independent variable and one or more independent variables. So, we have a dependent variable and we have an independent; we have one or more independent variables and how the dependent variable relates to one or more of these variables taken at a time or taken together.

(Refer Slide Time: 15:47)

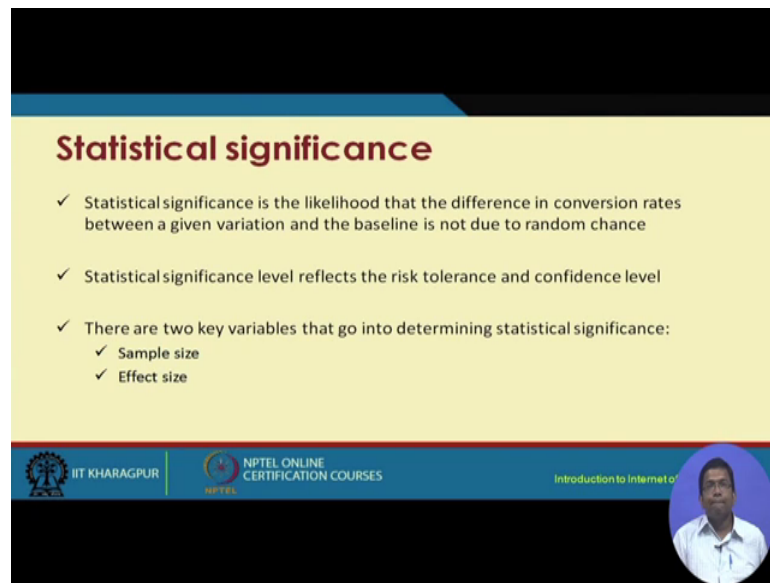So, regression analysis estimates the conditional expectation of the dependent variable given the independent variables. The estimation target is a function of the independent variables called the regression function. It characterizes the variation of the independent variable around the regression function which can be described by a probability distribution. So, regression analysis is helpful in different ways: it can be used to understand how the independent variables are related to the dependent variable one at a time or taken together.

Statistical significance is important. It basically measures the likelihood that the difference in conversion rates between given variation and the baseline is not due to any random chance. So, statistically how much the results are significant is something that has to be measured. So, statistical significance basically reflects the risk, tolerance and the confidence level. So, how much is the confidence on the results that are obtained.

So, this is measure through statistical significance. So, typically there are two key variables that are required for determining statistical significance: one is the sample size the other one is the effect size.
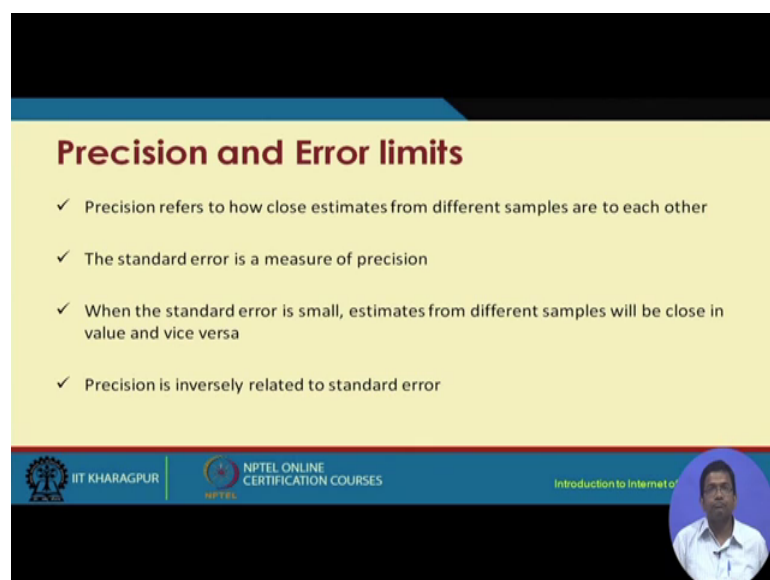
(Refer Slide Time: 17:07)

The sample size refers to the sample size of the experiment. The larger the sample size is the more confident one can be on the result of the experiment. And the effect size is just the standardized mean difference between the two groups. So, if a particular experiment is replicated the different effect size estimates from each study can easily be combined to give an overall best estimate of the effect size.
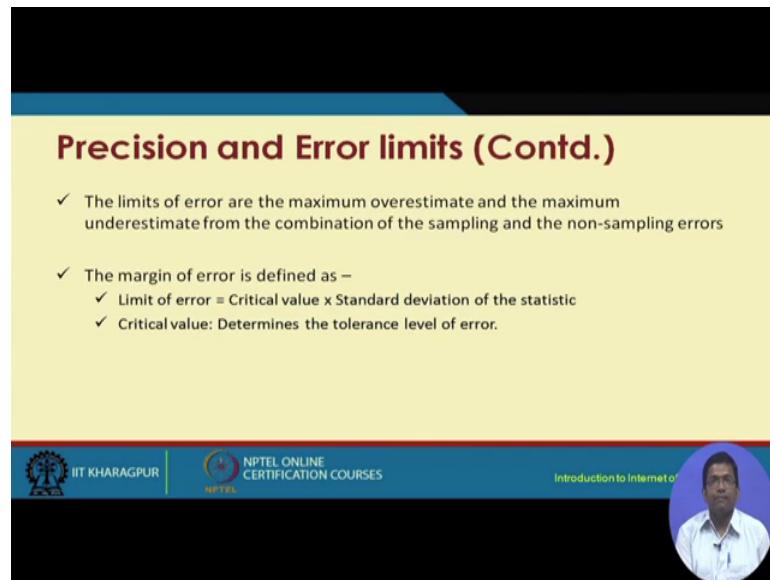
(Refer Slide Time: 17:38)

Precision and error limits are important. So, precision basically concerns how close the estimates are from the different samples to each other. The standard error is a measure of

the precision, when the standard error is small the estimates from the different samples will be closed in value and vice versa. So, precision is inversely related to the standard error.
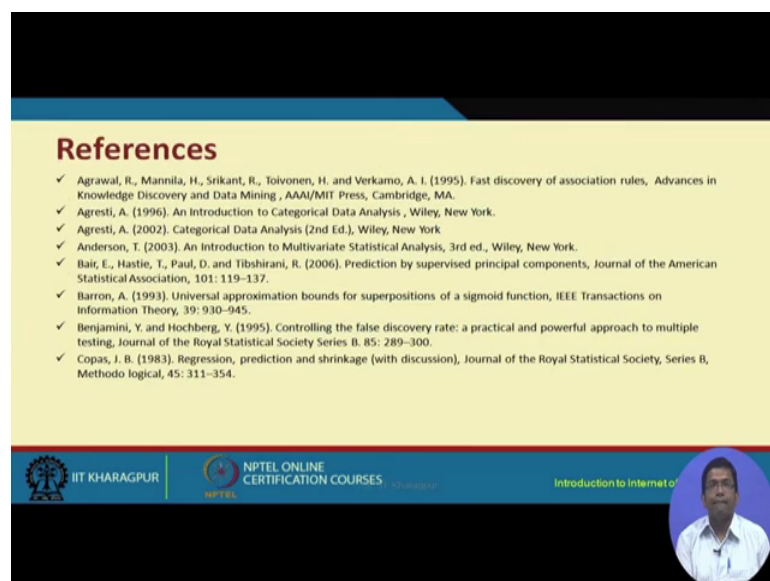
(Refer Slide Time: 18:03)

So, this precision an error become hand in hand, the limits of the error and the overestimate and the underestimate are taken into consideration while considering the error limits.

(Refer Slide Time: 18:19)

So, as I was mentioning at the outset that there are different statistical tools that additional statistical tools like correlation analysis, regression analysis, analysis of variance can be used in order to understand to how to get insight on the data that is obtained that is collected. But these are the basic analytic methods.

And what we have not discussed over here and we have intentionally confined ourselves to not discussing things like how text can be analyzed textual data or how video data can be analyzed and so on. So, that requires video data images can be analyzed and so on; so different other types of data can be analyzed. So, this we have intentionally not discussed, because that requires specialized training in text processing, video processing, image processing, and so on. And we do not want to get into the depth of those types of analytics.

So, these are the differences. So, with this we come to an end. And as I was telling mentioning before data handling, data analytics are very crucial in the context of IoT because lot of data gets generated in the IoT domain. And this data not only have to be analyze, but prior to analyzing they have to be handled. They have to be handled using technologies such as cloud, we have to be handled with technology such as (Refer Time: 19:49) and so on.

And once they are handled that means, the data have been stored, they have been cleaned and stored and so on then they have to be analyzed. For analysis we have these different statistical methods, we have different other methods based in machine learning, image processing, video processing, text processing and so on. So, those are advance methods which we do not cover in this particular lecture, in this particular course.

Thank you.