Statistical Learning for Reliability Analysis

Prof. Monalisa Sarma
Subir Chowdhury School of Quality and Reliability
Indian Institute of Technology - Kharagpur

Lecture – 38
Correlation Analysis (Part - I)

Hello guys. So, today we will start a new topic while before starting the topic, let us discuss few things like AMC annual maintenance contract I think all of you know about this right what is this? We have heard of this term annual maintenance contract for a water filter and annual maintenance contract for the washing machine for AC. So, it is something all of us we use it right in our day to day life we have to deal with is AMC. What is this?

This is basically a maintenance strategy and we do this we use this annual maintenance contract so, as to first so, that our system does not fail our machine does not fail. So, now, when the system becomes more and more complex, this annual maintenance contract it becomes very difficult as well as very expensive and a new asset a new engineering discipline as EMRs like it is called prognostic and help management.

To deal with such complex machine to different sort of maintenance strategy by which we can like find out we can like improve the state of the health of the system, we can monitor the health of the system and by that we can improve the health of the system one technique of PHM is called remaining useful life estimation. So, it is called RUL estimation remaining useful life estimation. So, it is RUL is gaining wide acceptance in most complex system and many of the complex system in RUL basically, it is totally data driven methodology.

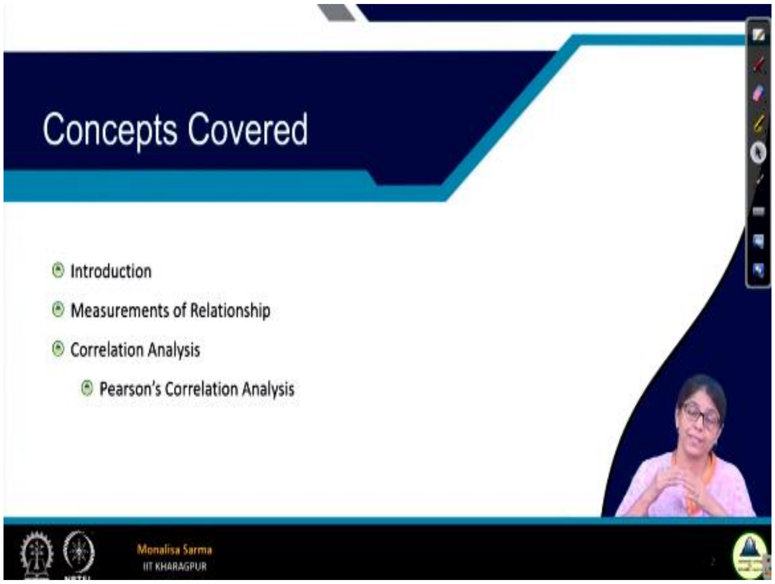
What we do here is that we use different data. Now, this data is what data does is condition monitoring data. This condition we take this data continuously we take the continuous monitoring data and how do we collect it there are different ways of collecting this data maybe we can collect this data by if we have some what to say sensors added in included in different equipment's in a complex system we have sensors in different, different components different sub components level.

And then we collect this data and from this data we try to estimate the remaining useful life. Now, how do we do that basically, we have lots and lots of data now, okay sensor data let us dig into sensor data, we have many sensors for a complex system say around 20 sensors 50 sensors and we have data from all the sensor this data may be ranging of different pressure variable temperature variable different capacitance value different resistance values different power converters, different power convert as well as a different rate how it degrades.

So, essentially what we need to do is that now, when we have this data, we need to analyze this data analyzing means there can be different things, we may want to find out is there any relationship between 2 different data 2 different variables of data like so, we may be interested in finding given a data can we predict some other data so in today's topic and income in fact, I will from this today's class onwards, I will be mainly dealing with I will be mainly talking with techniques for this data analysis.

So, this is what I the example what I have given is totally from the reliability perspective. So, now this data analyses technique, which I will be discussing, it is not only for not only applicable for what to say reliability is the studies it is equally applicable for any data analyses strategy, data analysis, wherever data analysis is necessary, it is this methods are applicable to all such situations.

(Refer Slide Time: 03:59)



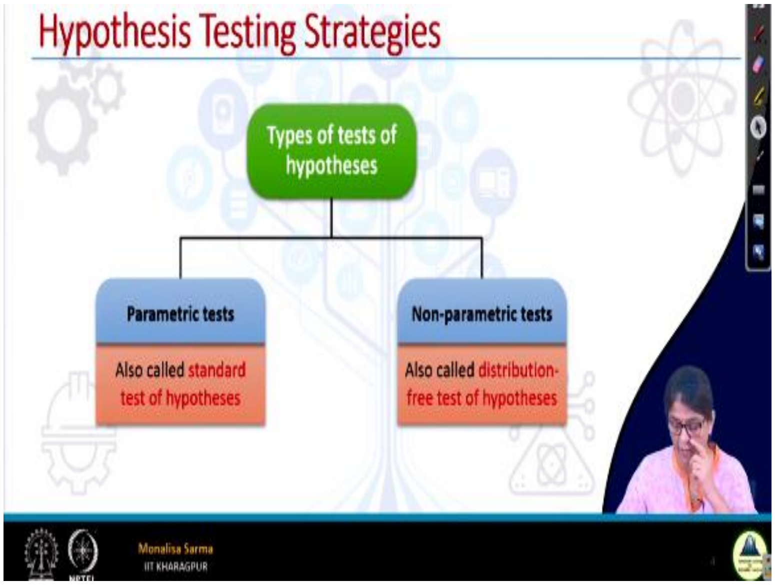
Concepts Covered

- Introduction
- Measurements of Relationship
- Correlation Analysis
  - Pearson's Correlation Analysis

So, now, coming to our discussion on data analysis move towards thing which we will learn here today is relational analysis, given the 2 variables of a data like when I talk about variables of data, I think, you know, because what are the variables of data we talked in my in my first lecture, first lecture series of this course. So, when if there are 2 variables of a data lake say temperature and pressure, so, now then, is there any relationship between these 2 data?

So, that is what we need to find out. Now, this relationship when we talk about this relationship, is there any way to measure this relationship? So, how weak the relationship safe is how strong the relationship is there some way to measure this relationship? If yes, then how that we will be discussing in this lecture. Now next, one of the methods of relationship analysis is correlation analysis. So under the banner of correlation analysis the different techniques one such technique is Pearson correlation analysis and today's lecture will be discussing Pearson correlation analysis.

(Refer Slide Time: 05:11)



Hypothesis Testing Strategies

Types of tests of hypotheses

Parametric tests

Also called standard test of hypotheses

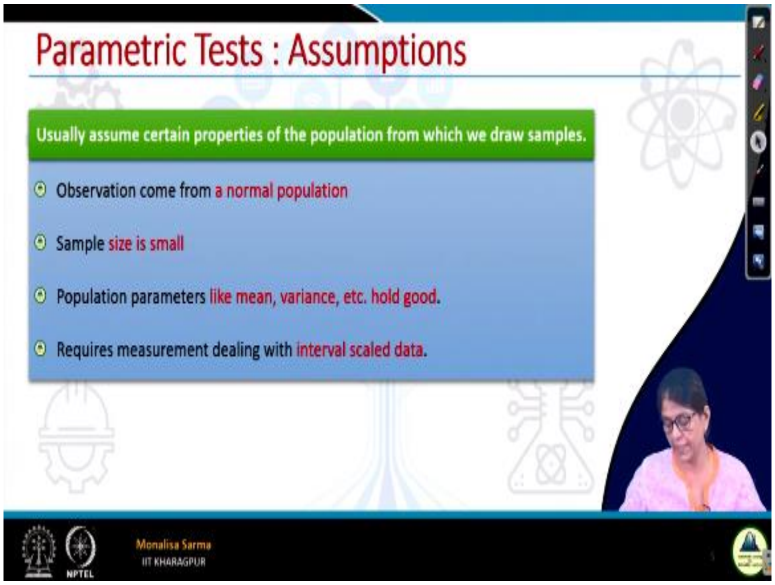
Non-parametric tests

Also called distribution-free test of hypotheses

So, now before going to relation analysis first take a quick recap of hypothesis test which we have done in a last couple of classes. So, the hypothesis test which we have done it is for that whatever we have assumed populations we have taken the populations from the we have taken a sample from the sample we have tried to find out the infer something about the population. And though for all these cases what we have talked about that is always we have assumed something about our population.

This type of hypothesis test is called parametric tests, so, it is also called standard test of hypothesis. And there's another one test which is called nonparametric tests. This is called the distribution free test of hypothesis we do not assume anything about the parent population here.

(Refer Slide Time: 06:02)



Parametric Tests : Assumptions

Usually assume certain properties of the population from which we draw samples.

- Observation come from a normal population
- Sample size is small
- Population parameters like mean, variance, etc. hold good.
- Requires measurement dealing with interval scaled data.

Now, so, this parametric test what we have used in the hypothesis testing till now, what we have done is parametric test parametric hypothesis testing that means, here we have certain assumptions about the populations what that we have considered like the populations has to be it has to come from a normal populations, I mean the parent population has to be normal populations, and the sample size is reasonable it is not very big populations, it is a reasonable size.

In fact, for a small sample if the population is totally normal, then small sample size will also so do and in the parametric cases, what we usually try to do is that we usually try to infer about the population mean and variance from the sample mean and variance from the sample statistics, we try to infer the populations mean and variance. So, population parameters mean and variance this sort of things holds good when we are talking about the parametric test.

Then the fourth thing is that it requires measurement dealing with interval scale data interval scale data actually, this is I did not mention it in my previous classes, as you mean that you guys already know about this. So, at this point, I think I will talk about it if you know it is very good, it will be a quick repair recap if you do not know that it is better than better means now you can learn actually, so, when I talk often variable and the first class have discussed variables there are 2 types of variables remember.

One is quantitative variable and one is qualitative variable quantitative variables basically is the numeric variable and another is qualitative variable. So, numeric variables basically there are 2 types of numeric data one is we call interval scale data and ratio data basically numeric variables only. So, an interval scale data and ratio scale data there are 2 types of numeric variables. So, there are only one difference is that difference is that an interval scale data 0 has meaning and a ratio scale data 0 has no value.

An interval scale data basically we call it interval scale data when we can measure the data in continuum within an interval, like set temperature ranging from - 10 degrees centigrade to + 10 degrees centigrade that is one interval another interval may be from 10 degrees centigrade to 30 degrees centigrade another may be 30 to 50 degrees centigrade. So, here when I have taken the range from - 10 degrees centigrade to + 10 degree here in between 0 is also there 0 has a value here. So, this is interval scale data.

But if at the same time if I measure the data in Kelvin instead of measuring in Celsius or Fahrenheit if I measure the data and Kelvin, so, I cannot take this range from - 10 to + 10 because 0 degree Kelvin does no value. Similarly, if you consider height, zero height has no value has no meaning. So, height cannot be interval scaled data it will be a ratio scale. So, like weight it will be a ratio scale so, this is both are numeric variables only with this slight difference on meaning of 0 in interval scale data there it is mean 0 has its own significance.

In ratio scale data 0 does not have any significance and this is for quantitative variable that is numeric variables. Now, for qualitative variables we also call this categorical variable I think you remember that so, categorical variable there are 2 types one is ordinal and another is nominal. So, what is nominal data nominal variables are those variables which we can divide into different groups and there is no intrinsic ordering among those.

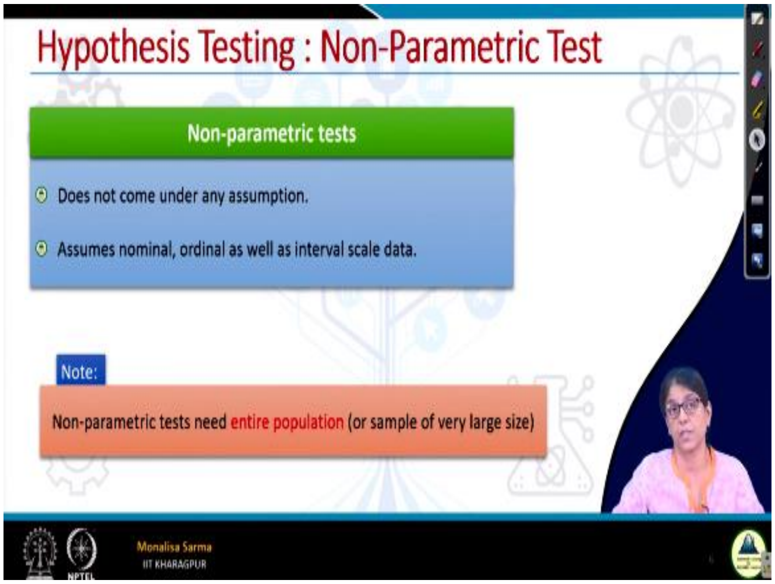
There is no order we can divide this into different categories, but then there is no ordering among these groups like you can take an example like PIN code PIN code of different places in a state. So, PIN code if we consider in a district maybe the PIN code of different places in a district if we take this different PIN codes different PIN codes indicates different places, but is there is no ordering among them. So, that is a nominal data.

Like again the nominal data one example I suddenly remembered. So now suppose I divide the population into different groups one group who is COVID vaccinated another group who is not COVID vaccinated. So, they are 2 different groups. So, they are nominal groups they are nominal data since there is no ordering I cannot say this is this is better than this is based on something that nothing not like that I cannot order them this is the has the highest order this is the lowest order nothing like that.

So, now, this nominal data if it has only 2 groups if it has a 2 group we; call it as dichotomous variable. So, you have had 2 groups let us suppose we consider gender if we consider gender has this to group male and female then it is a dichotomous variable dichotomous type of variable, it is basically a nominal variable but we can since it is this 2 group we can call it as a dichotomous variable. Similarly, now come to ordinal data ordinal data similarly, if we can divide this data into different groups, but there is intrinsic ordering among them.

Like the size of a garment size of a garment is very much an ordinal data size of a garment is triple x double x x medium small this is the ordering among them. So, then again policies of the government policies of the government may be very good, good, bad, worse. So, there is an ordering among them. So, these are called ordinal data. So, when we talk about parameter test so, parameter test it deals with an interval scale data.

(Refer Slide Time: 11:20)



Hypothesis Testing : Non-Parametric Test

Non-parametric tests

- Does not come under any assumption.
- Assumes nominal, ordinal as well as interval scale data.

Note:

Non-parametric tests need entire population (or sample of very large size)

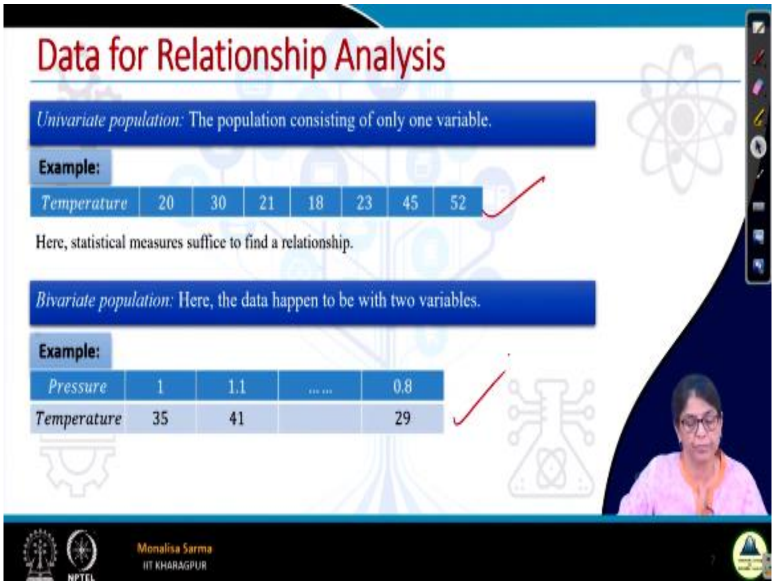
So, now nonparametric test, nonparametric test is does not come under any assumption and means parent population can be anything we do not have to assume anything about parent populations and it assumes it works with nominal data ordinal data interval scale data ratio data everything. And but the con the con of nonparametric is that it needs the entire population or a sample of a very large size.

So, now, the question comes is that at least is when I started with relational analysis, I wanted to talk about relational analysis why suddenly I started talking about parametric nonparametric and all this stuff? Basically to highlight the fact that till now, what we have discussed the different methods that have discussed is totally related to nonparametric tests all the methods we have for statistical inferences, what we have discussed all our nonparametric types of tests.

Now, from now onwards like when I will be discussing today, correlation analysis, there are some methods are parametric some methods are nonparametric. So, that is why I thought of bringing this into the picture. Now, when I will be discussing a parametric method to let you know that this is parametric and when this was nonparametric, I will let you know this is nonparametric. Anyway, I do not need to say also it will be obvious you will be able to find out okay this is a parameter method this is a non parameter method you also will be able to say.

When because you know what is the requirement for the parametric method what are the requirements for nonparametric method mainly from the data perspective when we talk of nonparametric method we can assume nominal data ordinal data any type of data.

(Refer Slide Time: 12:58)



Data for Relationship Analysis

Univariate population: The population consisting of only one variable.

Example:

Temperature	20	30	21	18	23	45	52

Here, statistical measures suffice to find a relationship.

Bivariate population: Here, the data happen to be with two variables.

Example:

Pressure	1	1.1	……	0.8
Temperature	35	41		29

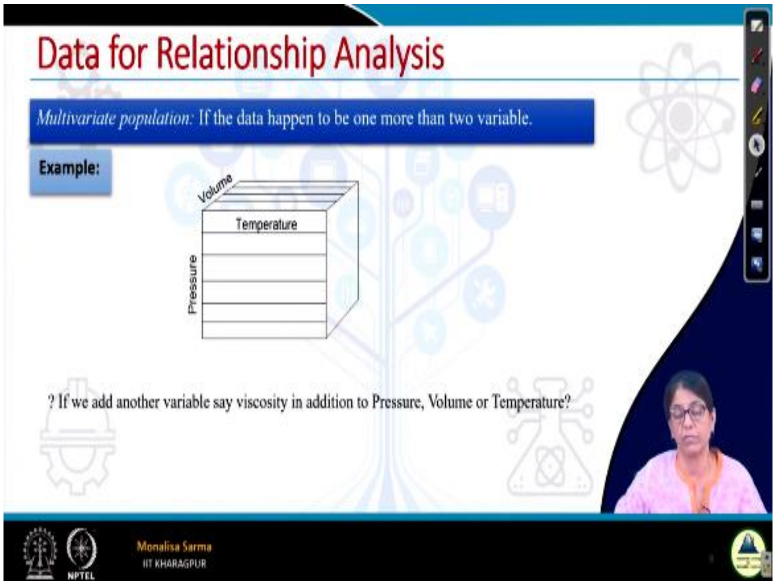
Monalisa Sarma
IIT KHARAGPUR

So, now, come to again coming back to the relationship analysis when we talk about relations, usually we talk of relations between 2 or more variables, we do not talk about relations with oneself it a single variable, suppose any populations there is only one parameter of interest, then we do not have any relations, there is only one parameter of interest, we are interested on only one term then maybe other parameter but we are interested in only one parameter then we do not talk a relation in such case.

Example, suppose you take the first example where it is what it is given here. So, this example, so, maybe this we can see that maybe these are the maximum temperature that is recorded in the 7 days in a week and the maximum recorded temperature recorded at MV. So, this is just only one variable. So, we are not interested in relation of such data and this is such thing, data is a call such population are called univariate populations.

So, then bivariate population may be in a system, consider a system we are interested in finding a pressure and we are interested in data pressure data and temperature data. So then it is called a bivariate population.

(Refer Slide Time: 14:09)



Data for Relationship Analysis

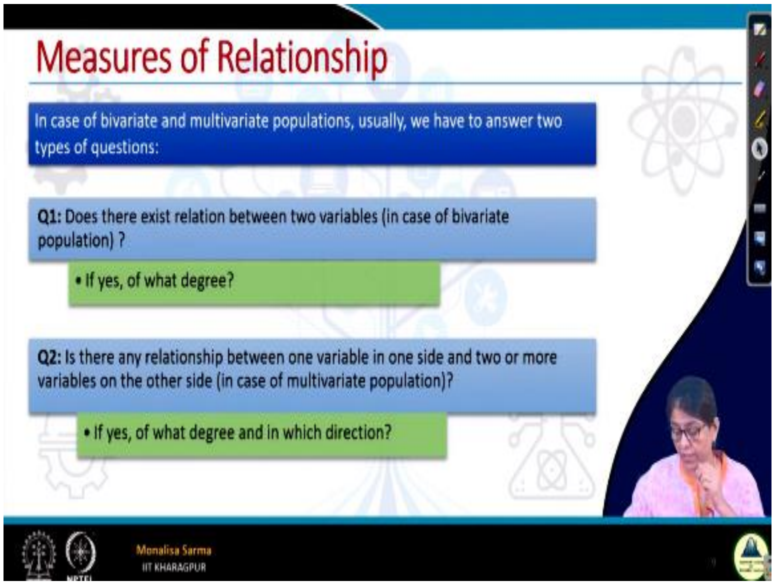
Multivariate population: If the data happen to be one more than two variable.

Example:

? If we add another variable say viscosity in addition to Pressure, Volume or Temperature?

Now, in for the same system, suppose we are interested in the volume as well. And suppose we add one more test is viscosity, then what happens it becomes a multivariate population. So when we talk about relation analyses, we talk about bivariate or multivariate, we do not talk about relation unless it is in case of univariate population.

(Refer Slide Time: 14:27)



Measures of Relationship

In case of bivariate and multivariate populations, usually, we have to answer two types of questions:

Q1: Does there exist relation between two variables (in case of bivariate population) ?

- If yes, of what degree?

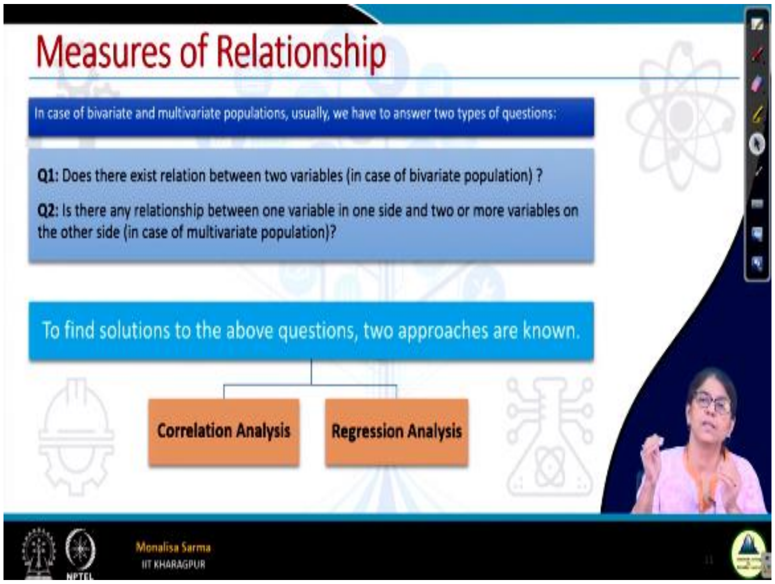
Q2: Is there any relationship between one variable in one side and two or more variables on the other side (in case of multivariate population)?

- If yes, of what degree and in which direction?

So, now, coming to this relationship, this relationship it seems very much like a categorical variable very much a qualitative variables. Now this qualitative variable, but really, is it a qualitative variable or we can measure it, if we can measure it, how we can measure it? So there are some questions which might come to our mind; like does there exists relationship between 2 variables in case of bivariate population, if yes, then of one degree, then when I talk about one degree that there comes some quantitative measure.

Similarly, is there any relationship between one variable in one side or 2 or more variables in the other side that is maybe I am talking about multivariate populations. So, if yes then what degree and in which directions.

(Refer Slide Time: 15:12)



Measures of Relationship

In case of bivariate and multivariate populations, usually, we have to answer two types of questions:

Q1: Does there exist relation between two variables (in case of bivariate population) ?

Q2: Is there any relationship between one variable in one side and two or more variables on the other side (in case of multivariate population)?

To find solutions to the above questions, two approaches are known.

Correlation Analysis

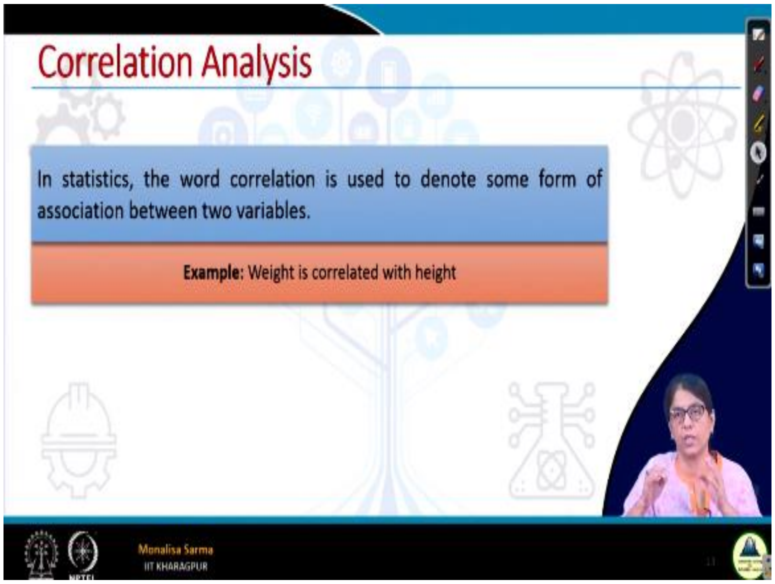
Regression Analysis

Monalisa Sarma
IIT KHARAGPUR

So, these are the problems we need a solution for this how we measure the relationship first is general relationship and then how do we measure this relationship. So, to find the solutions to there, the 2 approaches are known basically one is correlation analysis, one is regression analysis, correlation analysis and regression analysis both objectives are quite different. I will not talk regression at this stage I will talk and after 2 3 lectures maybe.

So, now, when I am interested in finding out the strength of the relationship, whether there is any correlation ship between the 2 variables and what is the strength of the correlation what is the strength of association is there any association between 2 relationship is any relationship between 2 variables and what is the strength of this association that is basically we will be learning that is the correlation analysis regression analysis is different.

(Refer Slide Time: 16:08)



Correlation Analysis

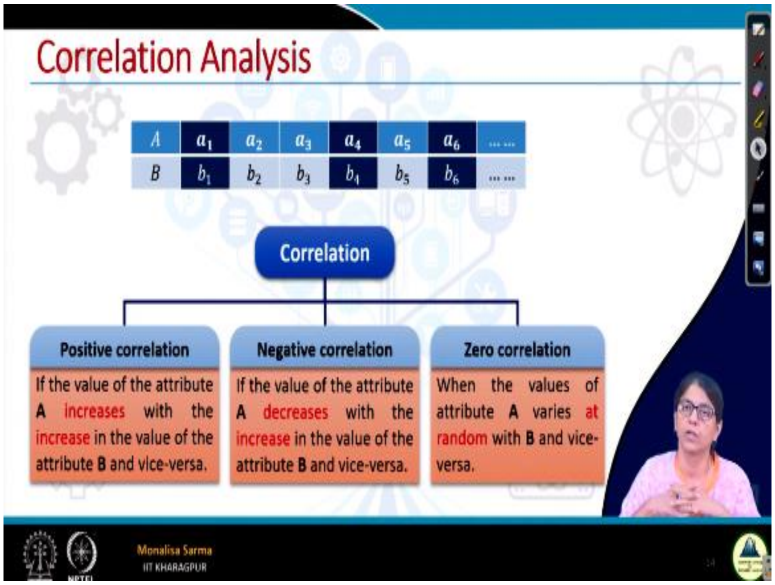
In statistics, the word correlation is used to denote some form of association between two variables.

Example: Weight is correlated with height

So, there in correlation in statistics the work correlation is used to denote some form of association between 2 variables, maybe weight is correlated with height, when height increases weight also increases maybe, or maybe absence from the class and score marks score in the test. So, when absence increases score decreases, the earlier the first example whatever even when height increases weight also increases maybe that is a positively positive correlation.

Based on when the next one what have told when I mean if you are absent from the class your test score decreases. So, one is increasing the other is decreasing this may be a negative correlation again maybe there are some factors where there is no correlation at all. So, we call it at zero correlations.

(Refer Slide Time: 16:54)



Correlation Analysis

A	a_1	a_2	a_3	a_4	a_5	a_6	... ...
B	b_1	b_2	b_3	b_4	b_5	b_6	... ...

Correlation

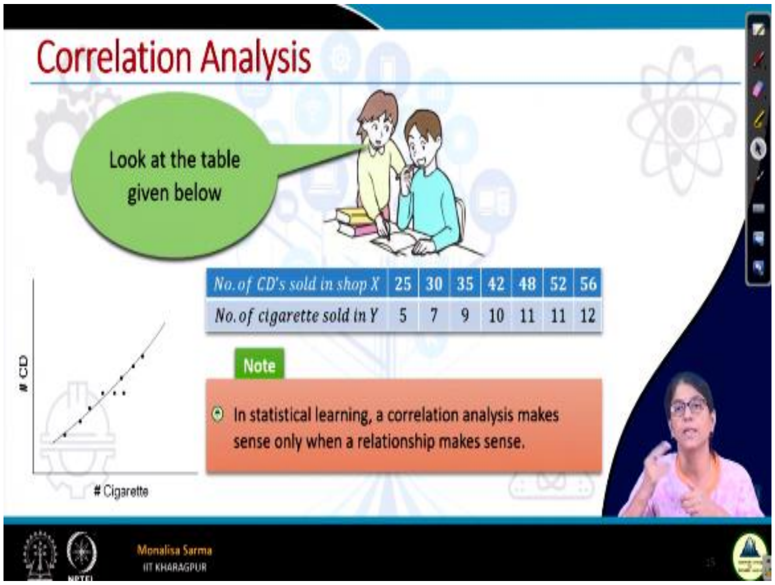
Positive correlation: If the value of the attribute A increases with the increase in the value of the attribute B and vice-versa.

Negative correlation: If the value of the attribute A decreases with the increase in the value of the attribute B and vice-versa.

Zero correlation: When the values of attribute A varies at random with B and vice-versa.

So, this is positive correlations if the value of the attribute A increases with the increase in the value of the attribute B and vice versa. Similarly, negative the value of A decreases with the increase of B and vice versa that is negative one is increasing other is decreasing other is decreasing one is next is increasing does this negative correlations and when the values of attribute A varies at random with B and vice versa, what is it? A is varying randomly it is not at all dependent on B when that is zero correlation.

(Refer Slide Time: 17:26)



Correlation Analysis

Look at the table given below

No. of CD's sold in shop X	25	30	35	42	48	52	56
No. of cigarette sold in Y	5	7	9	10	11	11	12

Note

- In statistical learning, a correlation analysis makes sense only when a relationship makes sense.

So, now, see here one example is given like we can see there are 2 columns basically I mean sorry, 2 rows one is showing the number of CDs sold in one shop X another is number of cigarettes sold in another shop Y. So, if we just see this data, one is talking about the CD selling of CD, another is talking about selling up cigarettes. So, we tried to find out the correlations and we found some correlation is there and we found this is how what is the strength of the correlation that also we found? But does it carry any meaning?

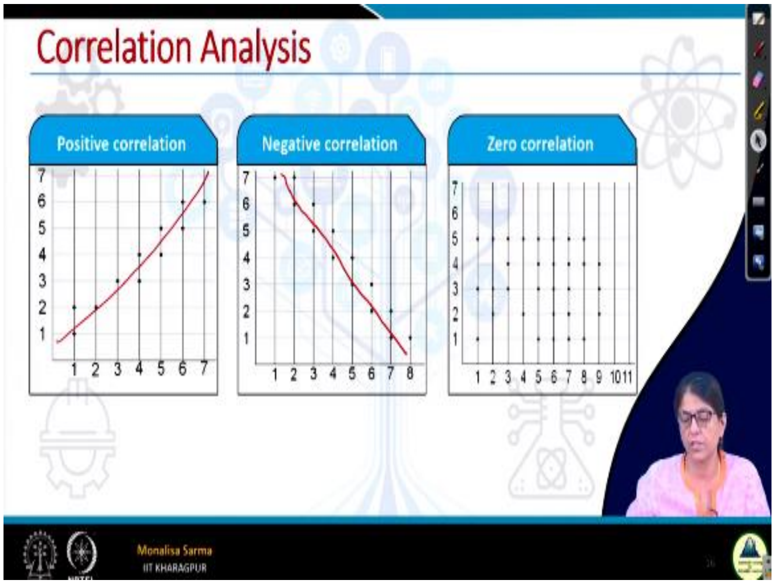
So, do we really need to find a correlation of distance. Now, there's the in statistical learning a correlation analysis makes sense when a relationship makes sense actually. So, just like that, we do not go on finding the correlation unless the relationship so make sense, then we go and find a correlation analysis. Now when I see this CD’s sold in one shop and cigarette sold in others, maybe in different places that carries no relations actually, that carries no meaning no sense.

But if at the same time if we see it in this way, if we watch lots of CD, that means you may smoke also more so if buying the CD and cigarette in a similar in a nearby vicinity, then maybe this has a correlation, so you are buying lots of CD means you will be spending too much time in front of your TV that means you may smoke also more so that way it has a relation. Now, one thing very important to note here is that when we were working on correlations analysis do not get caught confused with correlation with causality.

Causality means is cause and effect went because of something because of X Y is the effect cause because of because X is happening that is why Y is happening there is a direct cause and effect cause and effect is a different association altogether is a different association, it is called causality. So, we are not looking on this cause and effect analysis, we are just looking on this the correlation analysis cause and effect is just because of this; this is happening.

Now because of CD cigarette sell, because that definitely this is not the cause and effect of this causality analyses a different design of excellent techniques. We'll have to look into do that along with discussing in this lecture.

(Refer Slide Time: 19:45)



Correlation Analysis

Positive correlation

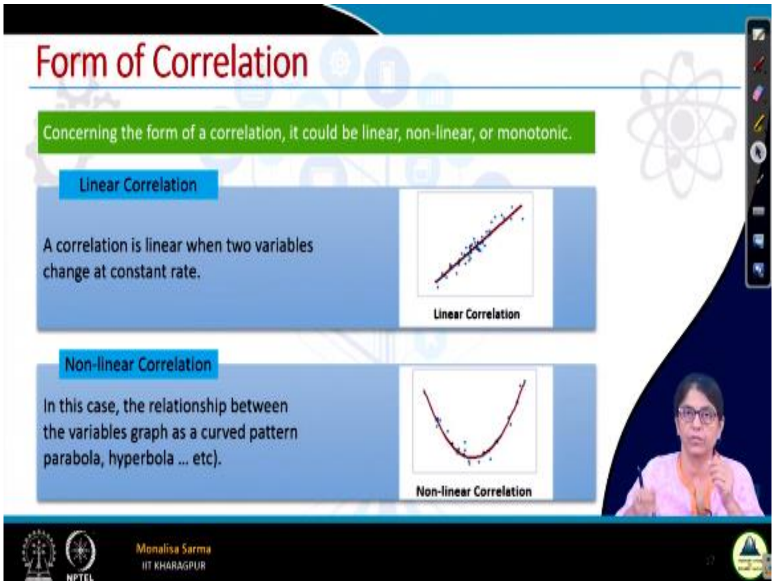
Negative correlation

Zero correlation

Monalisa Sarma
IIT KHARAGPUR

So now we can see the example of positive correlation here. So you see positive correlation. If I just draw a line here, then you will see the increase in one increases or the other So, similarly, this is negative correlation increasing one is decreases the other value if I take x and y coordinates, so x is one variable y as another variable okay. So, if y increases x decreases; if so, this is a negative correlation. Similarly, if this is the third figure there is no correlation at all the randomly it is ascending.

(Refer Slide Time: 20:24)



Form of Correlation

Concerning the form of a correlation, it could be linear, non-linear, or monotonic.

Linear Correlation

A correlation is linear when two variables change at constant rate.

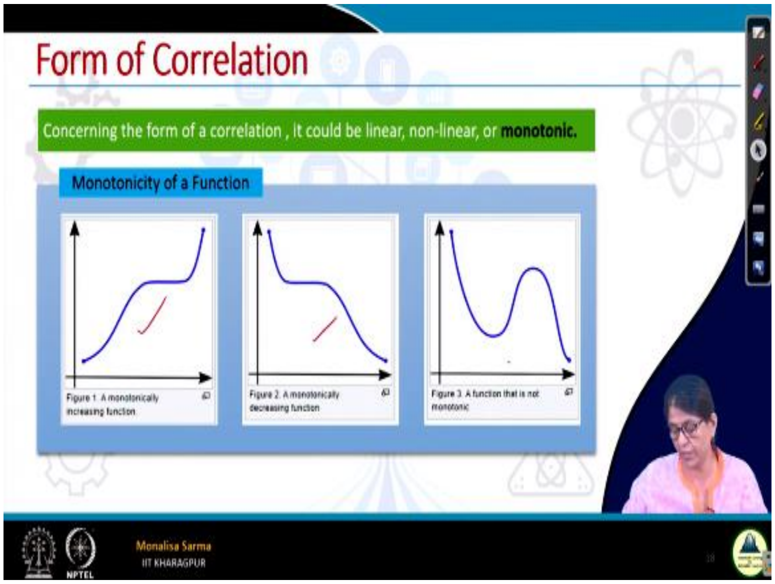
Non-linear Correlation

In this case, the relationship between the variables graph as a curved pattern parabola, hyperbola ... etc).

Monalisa Sarma
IIT KHARAGPUR

So, concerning the form of the correlations there can be different type of correlation it can be linear colinear or monotonic. So, if the relationship is linear between 2 variables then basically we can graph a straight line we can graph a straight line through the points then we call it as a linear when it is when the 2 variables change at a constant rate, then nonlinear correlations are less obvious and there is in the relationship between the variables it graphs as a cloud pattern like a parabola hyperbola etc. If we plot the dot it will graph as a curve pattern so, that is a nonlinear correlation.

(Refer Slide Time: 21:09)



Form of Correlation

Concerning the form of a correlation , it could be linear, non-linear, or **monotonic.**

Monotonicity of a Function

Figure 1. A monotonically increasing function

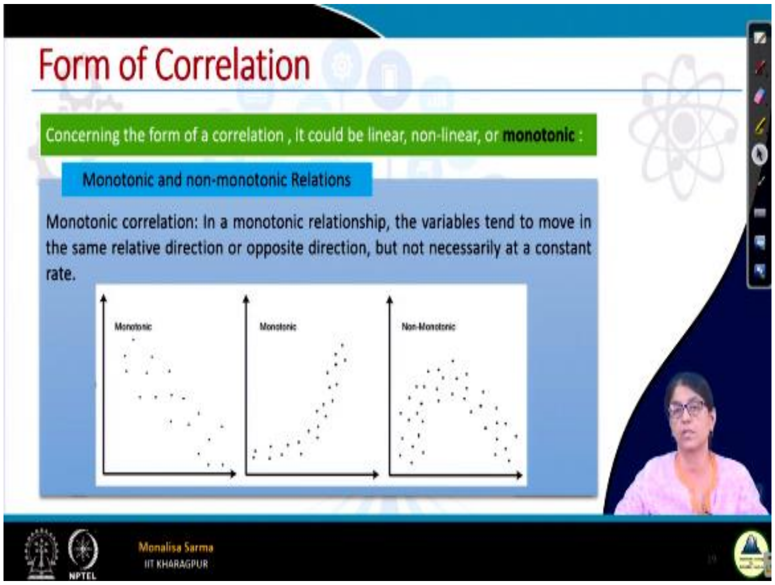
Figure 2. A monotonically decreasing function

Figure 3. A function that is not monotonic

Monalisa Sarma
IIT KHARAGPUR

And we have one more form of correlation and that is monotonic, monotonic also so, this is a monotonic function first one is a monotonically increasing this is monotonically decreasing and this is a non monotonic.

(Refer Slide Time: 21:20)



Form of Correlation

Concerning the form of a correlation , it could be linear, non-linear, or **monotonic** :

Monotonic and non-monotonic Relations

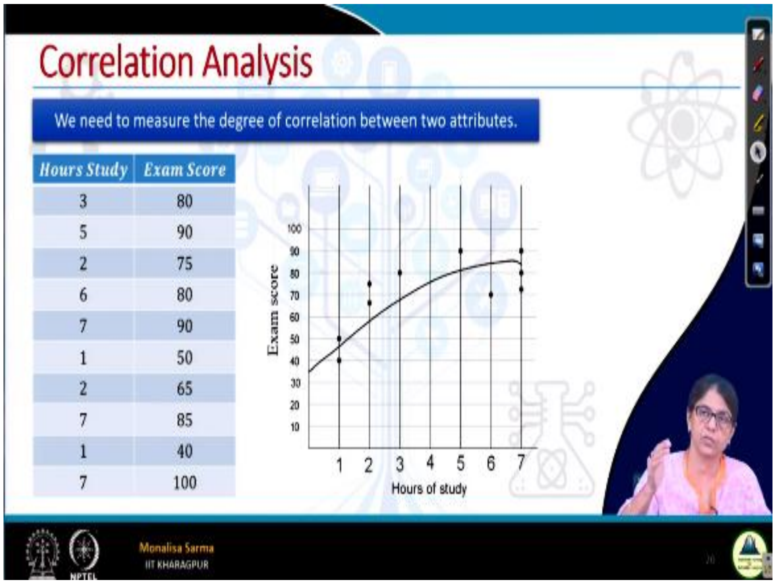
Monotonic correlation: In a monotonic relationship, the variables tend to move in the same relative direction or opposite direction, but not necessarily at a constant rate.

Monalisa Sarma
IIT KHARAGPUR

So, what is monotonic basically monotonic relationship the variables tend to move in the same relative direction or opposite direction like to linear relation which tend to move in the same direction or opposite direction similarly, in monotonic also same direction or opposite direction, but not necessarily at a constant rate and linear it moves at a constant rate either increasing or decreasing moves in a constant, but in case of monotonic it is not necessarily it moves in a constant rate.

So, a monotonic relation can be a really linear relationship, but a linear relationship will not be a monotonic relationship.

(Refer Slide Time: 21:52)



Correlation Analysis

We need to measure the degree of correlation between two attributes.

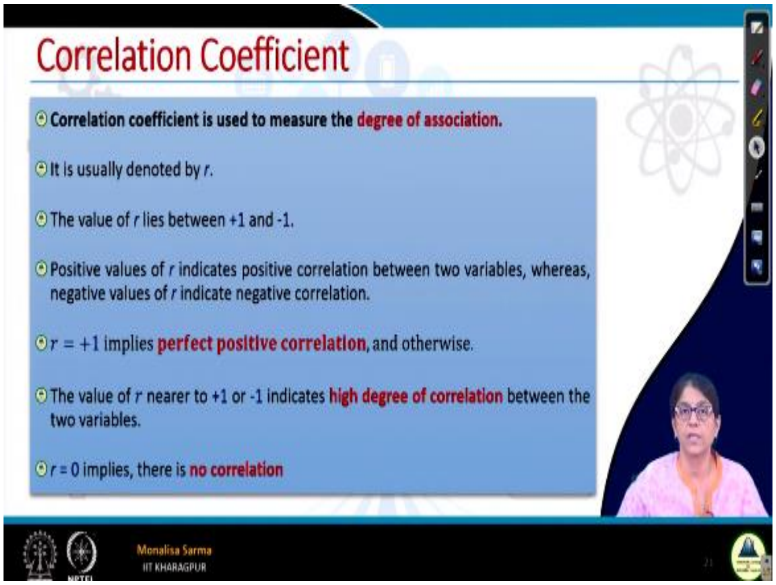
Hours Study	Exam Score
3	80
5	90
2	75
6	80
7	90
1	50
2	65
7	85
1	40
7	100

Monalisa Sarma
IIT KHARAGPUR

So, now, suppose we want to measure the relationship association degree of correlation between 2 attributes. Suppose the example given we have a data for few students, where students has put how many hours of study and then what is the exam score, so, is there any relationship between the hours of study they have put as well as the score they have received after putting this much hours of study. So, first, if we draw the graph first if we draw a scatterplot for this, for the data given we see that it is in positive correlations.

As one increases the other increases when we are putting more number of hours of study the other score is also increasing, because of course, all the points are not in the straight line, but definitely it is very near to the straight line some underline and some are just quite near to the straight line.

(Refer Slide Time: 22:43)



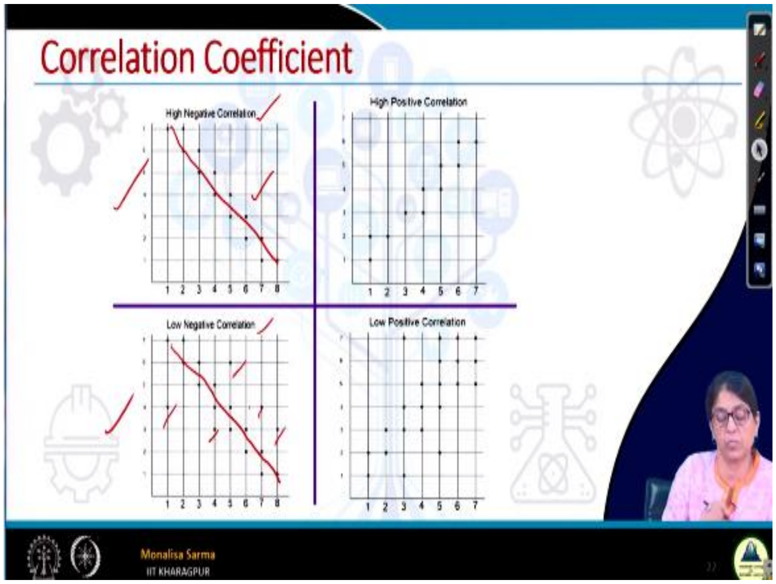
Correlation Coefficient

- Correlation coefficient is used to measure the **degree of association.**
- It is usually denoted by *r*.
- The value of *r* lies between +1 and -1.
- Positive values of *r* indicates positive correlation between two variables, whereas, negative values of *r* indicate negative correlation.
- *r* = +1 implies **perfect positive correlation**, and otherwise.
- The value of *r* nearer to +1 or -1 indicates **high degree of correlation** between the two variables.
- *r* = 0 implies, there is **no correlation**

So, now, how to find the strength of this correlation, how much this association is a strong association or it is a very weak association or there is no association at all. So, correlation coefficient is used to measure the degree of association. So, to find a degree of association we what we use is a? We use this correlation coefficient. So it is usually denoted by r correlation coefficient is denoted by r. So value of r lies between -1 to +1 so, if it is +1 then it is very highly correlated or near to +1 it is very highly correlated, 0 it is not there is no correlation.

And -1 also it is very highly correlated, but it is negatively correlated one increases the other decreases. So, the r = +1 implies perfect positive correlations. And the values of r near to +1 and -1 indicates high degree of correlation between the 2 variables, and r = 0 implies no correlation.

(Refer Slide Time: 23:42)



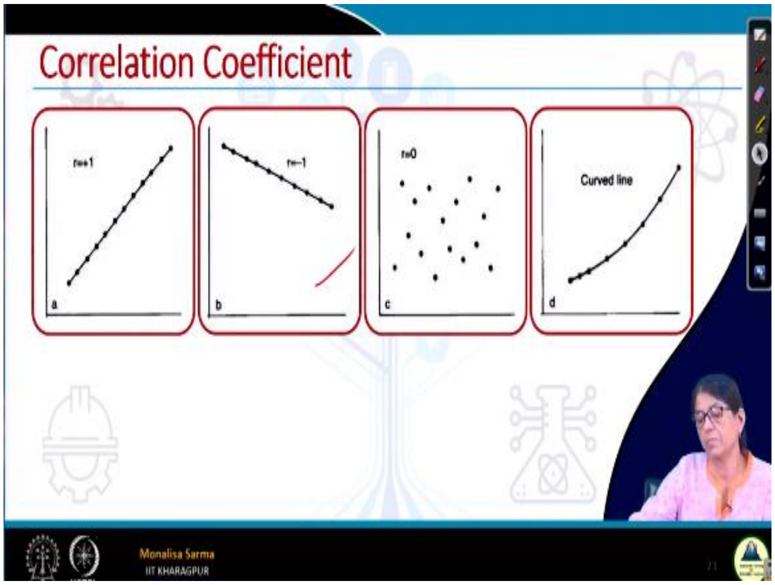
Correlation Coefficient

Monalisa Sarma
IIT KHARAGPUR

So now you see in this figure, the first one, this one, you see this figure and this figure, both have negative correlation. If I draw a line here, see, line here that runs basically closer to the points here also, I am drawing a line basically, which is closer to the point here, I am saying it is high negative correlation here, I am saying this low negative correlations, because you see the points here and discuss the points are dispersed. I am getting one point here, here, here, the points are dispersed.

It is the points are not all the points are not near to the straight line that I have drawn. But here you see all the points are nearer. So when I am trying to find the association, maybe I can maybe I am graphing a straight line to it because it is so the linear changes, but here the points are not very not very much disperse your points are very much dispersed. So it is a low negative correlation. Similarly for this example, high positive correlation low positive correlations.

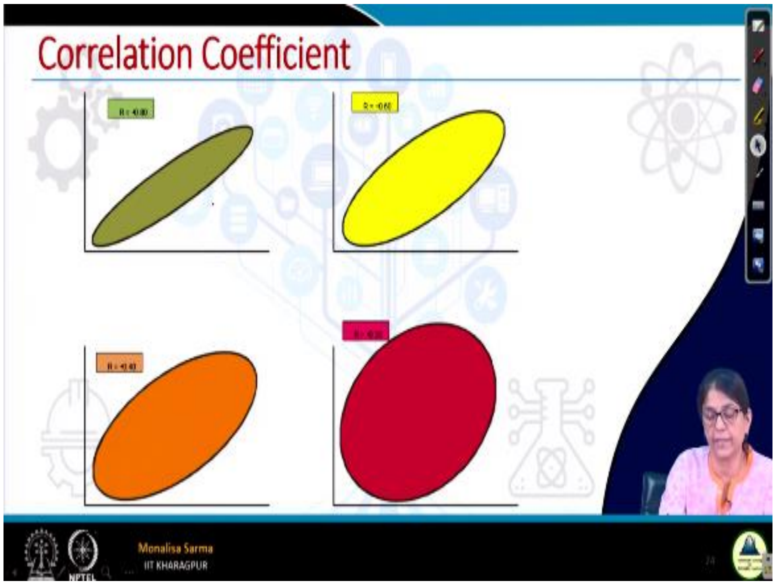
(Refer Slide Time: 24:35)



Correlation Coefficient

So, r is equals to + 1 something like that is also - 1 something like this figure. Then r is goes to 0. And this is a problem this may be monotonic fine.

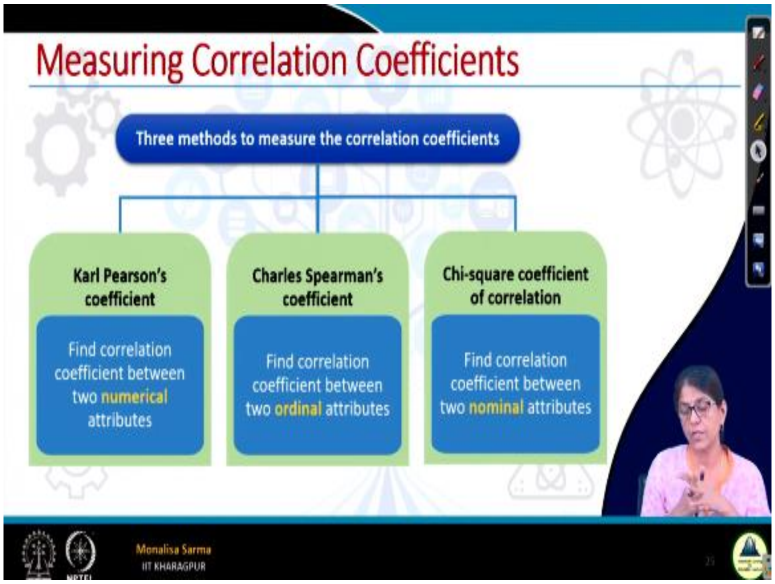
(Refer Slide Time: 24:50)



Correlation Coefficient

So, you see, have a figure for that. It says this figure we are I got this was 0.80 where the dispersion is there but not very high dispersion this person means for variance. But here in our last point we do a very high dispersion. Similarly is equals to 0.60 better, dispersion is greater than this but less than this is 0.30 again, this person more than more than 0.40 as well.

(Refer Slide Time: 25:20)



Measuring Correlation Coefficients

Three methods to measure the correlation coefficients

- Karl Pearson's coefficient: Find correlation coefficient between two numerical attributes
- Charles Spearman's coefficient: Find correlation coefficient between two ordinal attributes
- Chi-square coefficient of correlation: Find correlation coefficient between two nominal attributes

So, now, there are 3 different methods of there are not 3 different methods, there are many methods to measure the correlation coefficients, we will be discussing these 3 methods in this lecture not in this lecture in this course basically, today I will just discuss only Pearson's coefficients so, there we will be discussing three that is Pearson coefficients, Charles spearman’s coefficient and chi square coefficient of correlation.

So, Pearson coefficient it is used for numerical attributes Spearman we use for ordinal attributes remember ordinal attributes means data’s which are grouped into categories and which can be ranked. There is an intrinsic ordering among the categories and chi square provision its use for nominal attributes, there is no ranking among the data.

(Refer Slide Time: 26:04)

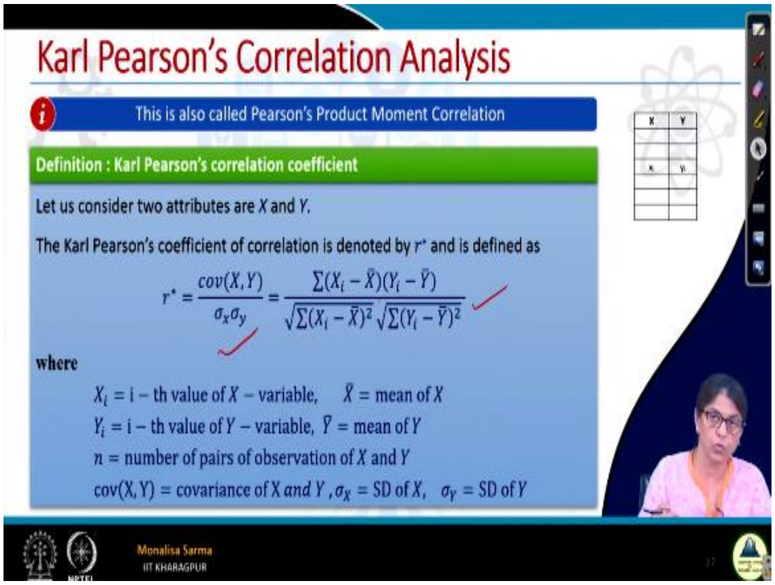
Pearson's Correlation Analysis

Statistical Learning for Reliability Analysis

Monalisa Sarma
IIT KHARAGPUR

So, in this lecture today, I will just discuss Pearson correlation analysis. So, Pearson correlation analysis it is as it is already mentioned here, that is between numerical attributes, we find try to try to find a correlation coefficient between numerical attributes. So, Pearson correlation coefficient hold is an analysis it is used to find is there any linear association is there any linear association between 2 variables.

(Refer Slide Time: 26:30)



Karl Pearson's Correlation Analysis

This is also called Pearson's Product Moment Correlation

Definition : Karl Pearson's correlation coefficient

Let us consider two attributes are X and Y.

The Karl Pearson's coefficient of correlation is denoted by r^* and is defined as

$$r^* = \frac{cov(X,Y)}{\sigma_x \sigma_y} = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2}\sqrt{\sum(Y_i - \bar{Y})^2}}$$

where

X_i = i – th value of X – variable, $\bar{X}$ = mean of X

Y_i = i – th value of Y – variable, $\bar{Y}$ = mean of Y

n = number of pairs of observation of X and Y

cov(X, Y) = covariance of X and Y , σ_X = SD of X, σ_Y = SD of Y

X	Y
X_i	Y_i

Monalisa Sarma
IIT KHARAGPUR

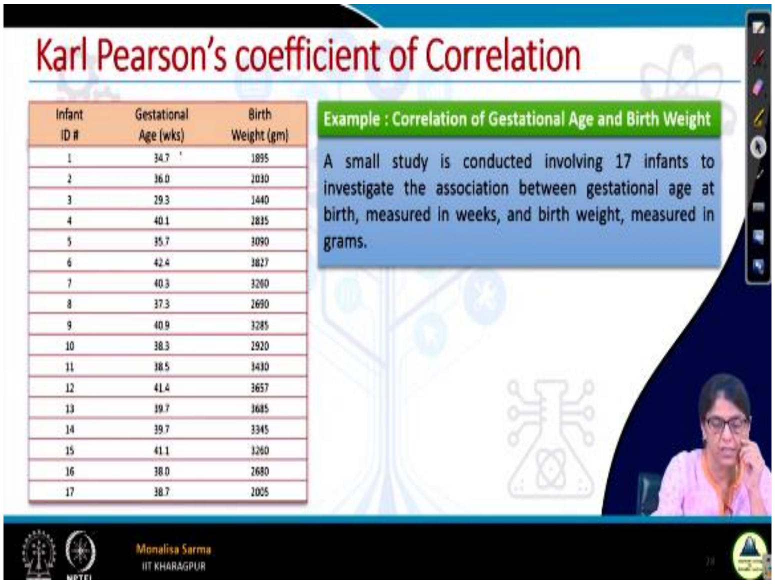
So, the data is it a data is numeric data then we use Pearson correlation analysis. So, now, how do we find out? How do we know can we use Pearson first when the set of data is given, we will try to first plot it maybe the scatterplot from the scatterplot. I will just get an idea. It is SOS is linear, then maybe I can use Pearson correlation analysis. Pearson correlation coefficient, as I told you, it tries to measure the correlation coefficient measure try to measure the strength of the correlation for 2 numeric variables.

And if the relationship is that linear, if there is a linear relationship maybe positive, maybe negative, but a relationship is linear than Pearson's correlation coefficient is applicable. So what is the formula for that we try and the formula for Karl Pearson's coefficient is this. This is the formula covariance of X and Y divided by standard deviation of x and standard deviation of y. Now, when I find out the covariance of X and Y my degrees when I will find a covariance of X and Y, what is the formula for covariance of X and Y?

This is the formula ∑ x i - x bar x y - y bar / degrees of freedom what will be my degrees of freedom total data and - 1 total data that is the n - 1. So, similarly, when I do σ x and σ y, so, σ x and σ y here there will be √ n - 1 n - 1 so, there will be one n - 1 to n - 1 n - 1 get cut. So, this is the formula for Pearson correlation coefficient essentially what I need to know X i is the ith value of this there are 2 variables X variable and Y variable.

So, different X i is the ith variable where i value goes from 1 to n and X bar is the mean of the all the X variables Y bar is the mean of the Y variables. So, putting these values together we will be able to find the correlation coefficient.

(Refer Slide Time: 28:19)



Karl Pearson's coefficient of Correlation

Infant ID #	Gestational Age (wks)	Birth Weight (gm)
1	34.7	1895
2	36.0	2030
3	29.3	1440
4	40.1	2835
5	35.7	3090
6	42.4	3827
7	40.3	3260
8	37.3	2690
9	40.9	3285
10	38.3	2920
11	38.5	3430
12	41.4	3657
13	39.7	3685
14	39.7	3345
15	41.1	3260
16	38.0	2680
17	38.7	2005

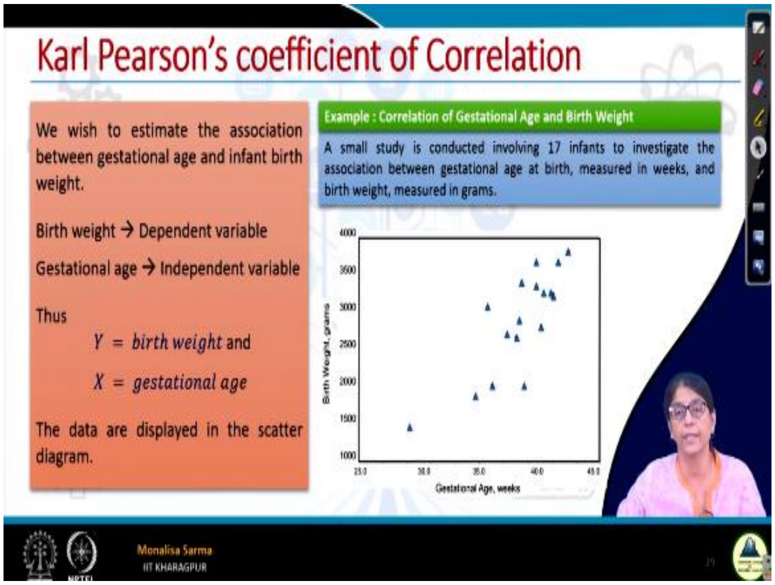
Example : Correlation of Gestational Age and Birth Weight

A small study is conducted involving 17 infants to investigate the association between gestational age at birth, measured in weeks, and birth weight, measured in grams.

Monalisa Sarma
IIT KHARAGPUR

So, here given an example suppose we want to conduct a small study involving 17 infants to investigate is there an association between the gestational age at birth measured in weeks and the birth weight. So, we want to find is any association between gestational age and what is the weight of the baby at the birth is there any association between that if we want to find out the ones and these are the values given so, this these are the values of gestational age which is given in weeks how many weeks and this is the birth weight is in grams.

(Refer Slide Time: 29:02)



Karl Pearson's coefficient of Correlation

We wish to estimate the association between gestational age and infant birth weight.

Birth weight → Dependent variable

Gestational age → Independent variable

Thus

Y = *birth weight* and

X = *gestational age*

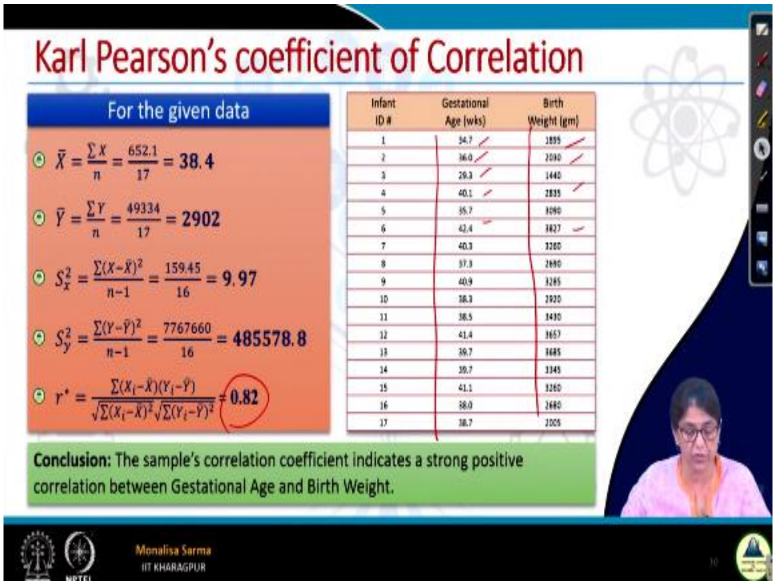
The data are displayed in the scatter diagram.

Example : Correlation of Gestational Age and Birth Weight

A small study is conducted involving 17 infants to investigate the association between gestational age at birth, measured in weeks, and birth weight, measured in grams.

So, we have done first we have done is scatterplot this is the one we got. So, we could see there is a linear association. So, let us see not that we will use Pearson correlation coefficient to find out what is the strength of this association.

(Refer Slide Time: 29:17)



Karl Pearson's coefficient of Correlation

For the given data

- $\bar{X} = \frac{\sum X}{n} = \frac{652.1}{17} = 38.4$
- $\bar{Y} = \frac{\sum Y}{n} = \frac{49334}{17} = 2902$
- $S_x^2 = \frac{\sum(X-\bar{X})^2}{n-1} = \frac{159.45}{16} = 9.97$
- $S_y^2 = \frac{\sum(Y-\bar{Y})^2}{n-1} = \frac{7767660}{16} = 485578.8$
- $r^* = \frac{\sum(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum(X_i-\bar{X})^2}\sqrt{\sum(Y_i-\bar{Y})^2}} = \boxed{0.82}$

Infant ID #	Gestational Age (wks)	Birth Weight (gm)
1	34.7	1895
2	36.0	2030
3	29.3	1440
4	40.1	2835
5	35.7	3090
6	42.4	3827
7	40.3	3260
8	37.3	2690
9	40.9	3285
10	38.3	2920
11	38.5	3430
12	41.4	3657
13	39.7	3685
14	39.7	3345
15	41.1	3260
16	38.0	2680
17	38.7	2005

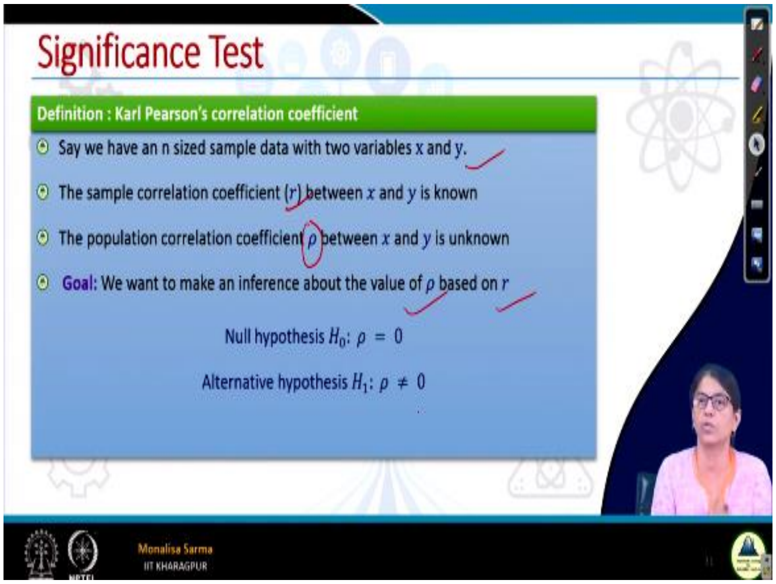
Conclusion: The sample's correlation coefficient indicates a strong positive correlation between Gestational Age and Birth Weight.

Monalisa Sarma
IIT KHARAGPUR

So, from this whatever values we will require X bar will require Y bar then for all X we will have to find out the this is simple just simple use of the formula just formula putting nothing else the values are given these are all X i values these are X 1 X 2 X 3 or X i values for till 17 these are all Y i values Y 1 Y 2 Y 3 and now we will have to find a mean of these we will have to find a mean of this mean of this will be X bar mean of this will be Y bar, just put it in a formula.

And at last what we got we got a value of 0.82. 0.82 is that means it is always highly correlated that means gestational age and birth weight is highly correlated. So, the sample correlation coefficient indicates a strong positive correlation between gestational age and birth weight because 0.84 is it is very much close to 1 it is it is a strong correlation strong positive correlation.

(Refer Slide Time: 30:14)



Significance Test

Definition : Karl Pearson's correlation coefficient

- Say we have an n sized sample data with two variables x and y.
- The sample correlation coefficient (r) between x and y is known
- The population correlation coefficient ρ between x and y is unknown
- **Goal:** We want to make an inference about the value of ρ based on r

Null hypothesis H_0: $\rho = 0$

Alternative hypothesis H_1: $\rho \neq 0$

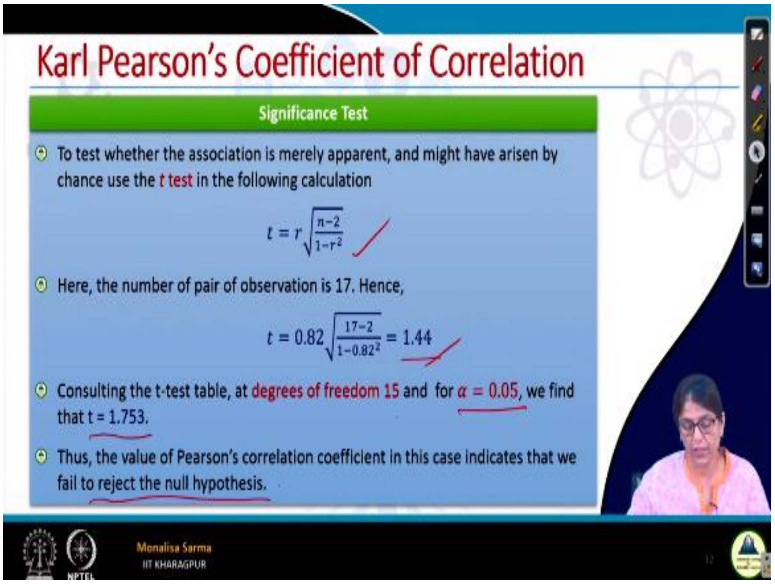
Now, the thing is that sometimes what happens because this we are doing bring it from a sample whatever it is, we are doing it from a sample we are not doing it for the whole data. So, when we are doing it from a sample maybe it happens that the sample that we have collected is by chance this coordination is shown here we got a very high positive correlation the sample that we have collected by chance by in simple Mayor by some coincidence, maybe we got this correlation, a high positive correlation.

Maybe it is not reflective of the actual populations, that is what we have done in hypothesis testing when we have tried to enforce those statistical inference similarly, for correlation analysis also we will have to find a significant stays will have to form the hypothesis will have to have the significance level same method everything same. So, here what is say we have n sized sample data with 2 variables x and y.

The sample correlation coefficient between x and y is r the sample correlation coefficient r that is this last value, what we got this 0.82 this is the r value from the sample we got. So, the population correlation coefficient is ρ between x and y is unknown population correlation coefficient is not known to us we know sample correlation coefficient. So, we want to make an inference about the value of ρ based on r like when we are doing hypothesis testing different way.

We wanted to infer the population mean from sample mean population variance from sample variance. Similarly, here we want to make an inference of the value of ρ based on r. So, my null hypothesis is ρ = 0 there is no correlation, alternate hypothesis is ρ ≠ 0, there is some correlation. There is a significant correlation.

(Refer Slide Time: 32:05)



Karl Pearson's Coefficient of Correlation

Significance Test

- To test whether the association is merely apparent, and might have arisen by chance use the *t* test in the following calculation

$$t = r\sqrt{\frac{n-2}{1-r^2}}$$

- Here, the number of pair of observation is 17. Hence,

$$t = 0.82\sqrt{\frac{17-2}{1-0.82^2}} = 1.44$$

- Consulting the t-test table, at degrees of freedom 15 and for $\alpha = 0.05$, we find that t = 1.753.
- Thus, the value of Pearson's correlation coefficient in this case indicates that we fail to reject the null hypothesis.

So, how do I do that to significant test is basically, in this case, we will do use the t test whatever t test formula we have used while doing statistical inference for mean of the populations. So, here also we are using t test but the representation of this t whatever representation we are using we are using a different representation, but actually we can from that we can get this from this we can get that. So, when it is not you do not take this why there we have used a different t when we use the t test we have used a different expression here we have using a different formula.

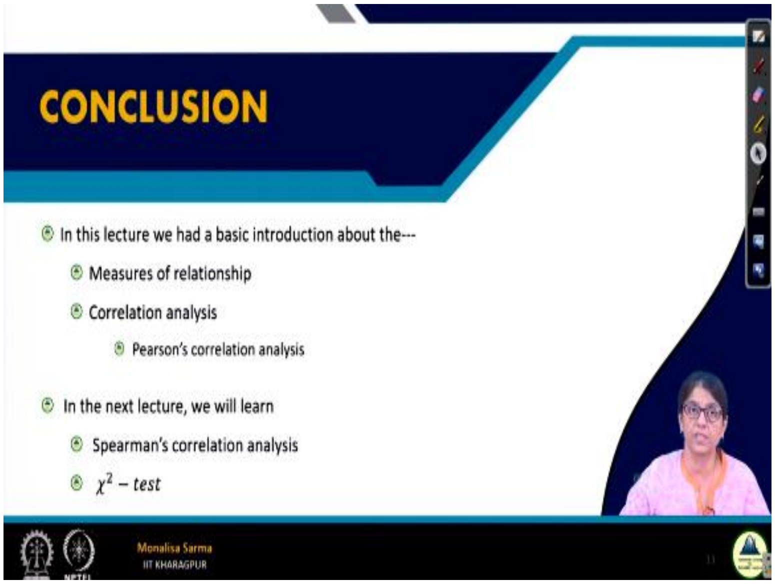
No, it is just a different representation. So, here we will be using this representation. What is that? Because here, we; need the t in terms of r that is the correlation coefficient. And there we need a t in terms of mean and standard deviation, the standard deviation of the sample. So, this is our formula for t. So, we will put everything here we know what is r what we have calculated n is the total sample size. So putting here so what we got we got a value of 1.44, suppose we have as we made significant level of 0.05.

So, for a significance level of point 05 degrees of freedom is 15 and - 2 because we are considering to between 2 variables within 2 sets of data each has n data, so n - 1 - 1. So, it is n - 2. So, for degrees of freedom 15 and alpha equals to 0.05 if we see the t table, t table, we have already seen lots of t table I think you already know that if we say the t table will get the value t = 1.753. That means our critical reason starts from 1.753 and a value greater than 1.753 it falls in a critical region it falls in rejection region.

So if it falls in the rejection region, and what happens then we reject the null hypothesis, is not it? We reject the null hypothesis and we accept the alternate hypothesis. Now, what we got we got a value of 1.44, 1.44 is very much in the acceptance region. So does the value of Pearson’s coefficient in this case indicate that we fail to reject the null hypothesis. See, here, we got a strong correlation coefficient, but actually in the population there is no correlation.

So that the data that we have got is totally purely by chance, so in this indicates that we fail to reject the null hypothesis and what is the null hypothesis? Null hypothesis is that there is no relation between the 2.

(Refer Slide Time: 34:37)



CONCLUSION

- In this lecture we had a basic introduction about the---
  - Measures of relationship
  - Correlation analysis
    - Pearson’s correlation analysis
- In the next lecture, we will learn
  - Spearman’s correlation analysis
  - $\chi^2 - test$

Monalisa Sarma
IIT KHARAGPUR

So, we end this lecture here. So in this lecture, what we have learned? We will learned how to measure the relationship and we have also seen Pearson’s correlation analysis. And in my next lecture I will discuss Spearman’s correlation analysis and chi square test.

(Refer Slide Time: 34:53)

REFERENCES

- The detail material related to this lecture can be found in

  The Elements of Statistical Learning, Data Mining, Inference, and Prediction (2nd Edn.), Trevor Hastie, Robert Tibshirani, Jerome Friedman, Springer, 2014.

Springer Series in Statistics
Trevor Hastie
Robert Tibshirani
Jerome Friedman
The Elements of Statistical Learning
Data Mining, Inference, and Prediction
Springer

Monalisa Sarma
IIT KHARAGPUR

So this is the reference. Thank you.