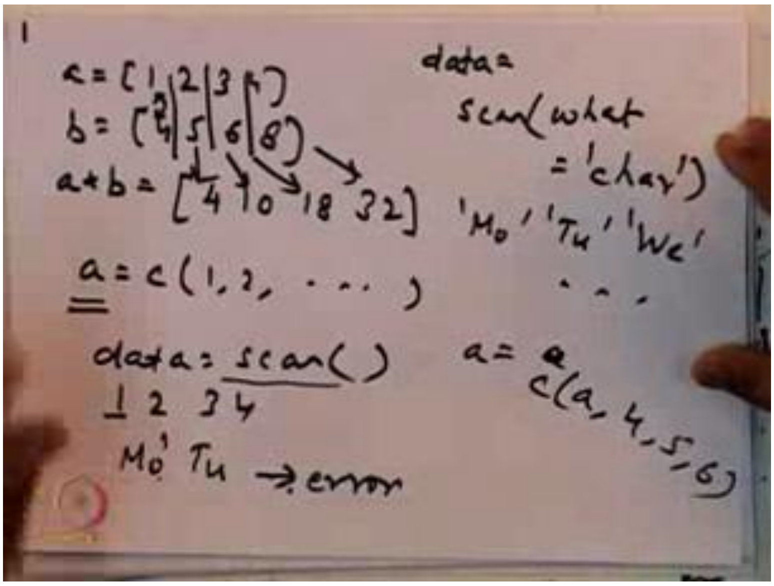
Introduction to Biostatistics

Prof. Shamik Sen
Department of Bioscience and Bioengineering
Indian Institute of Technology, Bombay

Lecture - 10
Correlation

Hello and welcome to today's lecture. I hope last week you got the opportunity to do some examples in R. In last class we had done; we had solved few examples in R showed you how you can create a vector; you can do basic scalar in addition, subtraction and other operations and how to do vector operations. So, one of the things to remember is when you define to vectors and you are doing these operations these operations get operated at an element wise level.

(Refer Slide Time: 00:49)



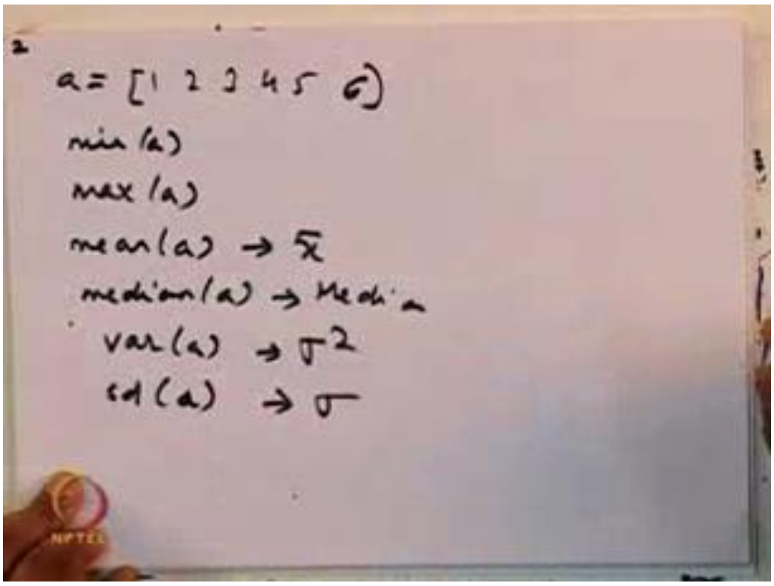
In other words, if I define as a vector which is 1, 2, 3, 4 and b as a vector which is 4, 5, 6, 8, then a star b will give me a vector which is 4, 10, 18 and 32 so on and so forth. So, basically you have an element wise operation a star b is operated at the element wise level and each product is given you as a separate vector.

We had then shown that you know how you use the scan. So, either you write a is equal to c of 1, 2, dot, dot, dot. So, this is the you know syntax for entering number and concatenating them into a single vector, but of course, this can be very laborious when you have big data. So, and you know repeatedly entering next to each other this can be a problem. So, one of the ways around, it used to use the function scan so you can write data is equal to scan and when you put note you know no brackets when you enter. So, you have the command prompt and where you can enter and these numbers get stored.

Most important thing to note is in this case, when you write the function scan by default the software assumes that these numbers are real. So, if I write Monday or Tuesday then immediately this will give to an error. So, the way around it is to write data is equal to scan and you have this additional term as what is equal to char. So, then tell the software then knows that you are essentially entering characters while you are entering these. So, then if you enter month but when you enter, you have to write it within quotes Monday, Tuesday, Wednesday, so on and so forth.

And then it will automatically take it you can easily add. So, you have an a vector a you can add let us say you can write a is equal to c of a comma 5, 4, 5, 6, so on and so forth. So, you can add these numbers either before or after the vector.

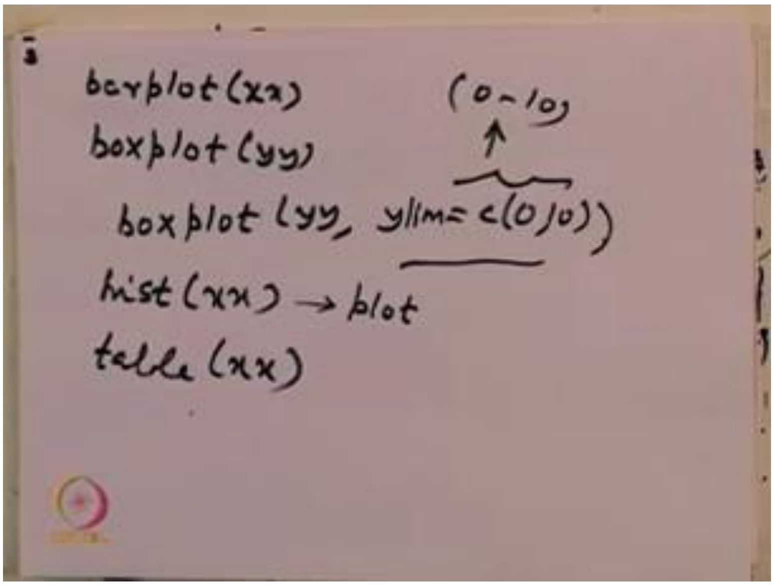
(Refer Slide Time: 02:56)



So, once you generate a vector let us say you have a vector a is equal to 1, 2, 3, 4, 5, 6, you can use these essential functions like min of a, max of a, mean of a, median of a, variance of a and s, d of a, to get variance sigma, square sigma this is just the median you get x bar here and min and max.

So, these functions will easily allow you to calculate numbers particularly when these vectors are b or the numbers are b then we briefly discussed about plotting.

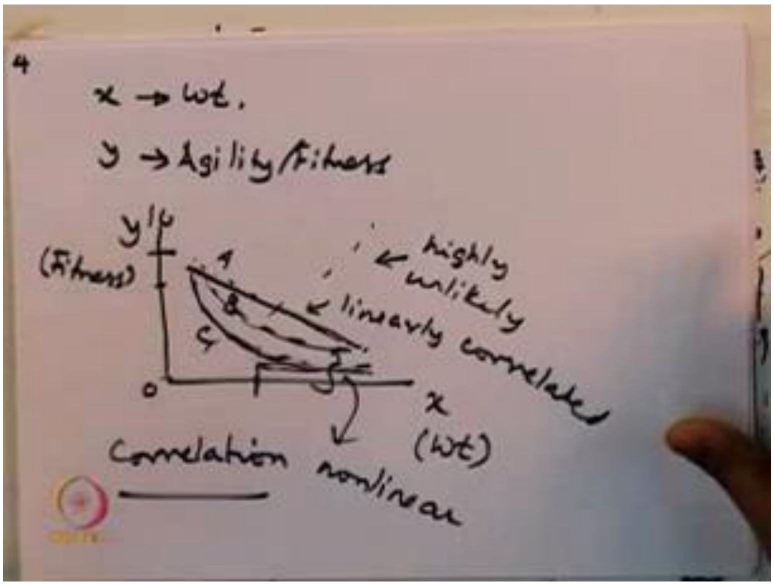
(Refer Slide Time: 03:46)



So, once you have a vector you can use let us say a box plot. So, if you use like bar plot or bar plot of let us say x x or box plot of x x y y, you will generate all these plots in the plotting function. So, you let us say I could have written box plot y y and then I could have written ylim is equal to c of 0 to 10. So, this would have said the y axis range. So, this is the y axis range from 0 to 10. So, this is how I would have entered my y axis range to be between 0 and 10 so on and so forth.

Histogram of x x or y y will give you the histogram or the frequency distribution, but it will also generate the plot. So, just to get the frequency distribution you can write table of x x. So, these are the basics. Now let us come to say in the generic case you just do not have values, but you have values where there are more than one metric when chosen.

(Refer Slide Time: 04:55)



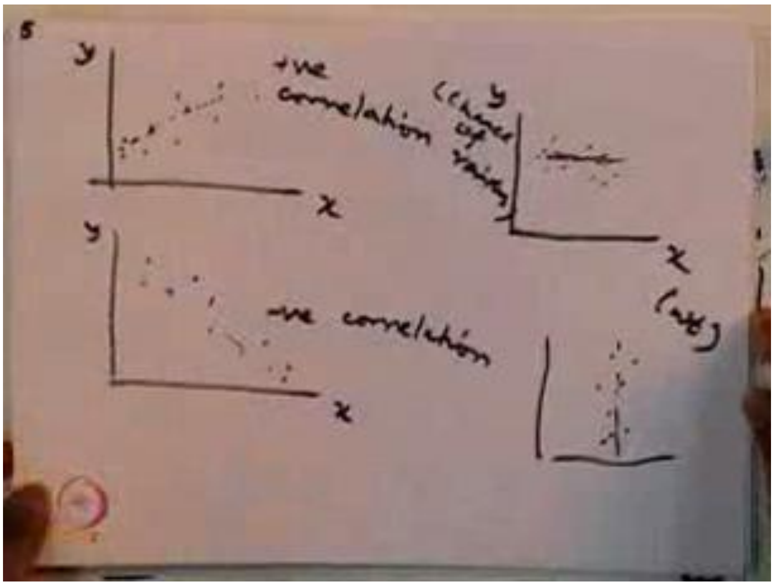
So, let us say in a class I want to correlate. So, you have 2 vectors you have chosen x and y of these x is let us say weight and y is agility or you know capability to run let us say how agile or fitness whatever you choose now logic would dictate you would expect in general that if I were to plot x and y if this is my x this is my y. So, x is my weight axis and y is my fitness axis and let us say I you know I normalize it with respect between a value between 0 and 10. So, you would expect that as weight will drop you would you can expect a curve like this you can expect a curve like this you can expect a curve like this, but it is highly unlikely that you will have a curve like this is highly unlikely from a physical point of view.

So, the object of this exercise is to correlate this to particular behavior and this is how is chosen in the principle of correlation how are they correlated. So, I can clearly see that in both let us say this curve a, this curve b and this curve c they are correlated. So, as per this curve a let us say they are saying that you see a strong correlation such that increase in weight gives rise to decrease in fitness in b this b or for that matter c this is much stronger. So, it says that even for small changes in weight initially there is a huge drop in the fitness of the person concerned.

But beyond a certain weight you have saturation. So, clearly you can see that depending on the nature of the data you might see these 2 curves can be linearly correlated or for these 2 curves this relationship in non-linear that is with linear increase if you are you know if your weight double will your fitness also reduced by half that is not so.

So, these principles are very useful for studying correlation and regression and let us see; how it is done. So, how do you know whether something is positively correlated or something is negatively correlated?

(Refer Slide Time: 07:23)

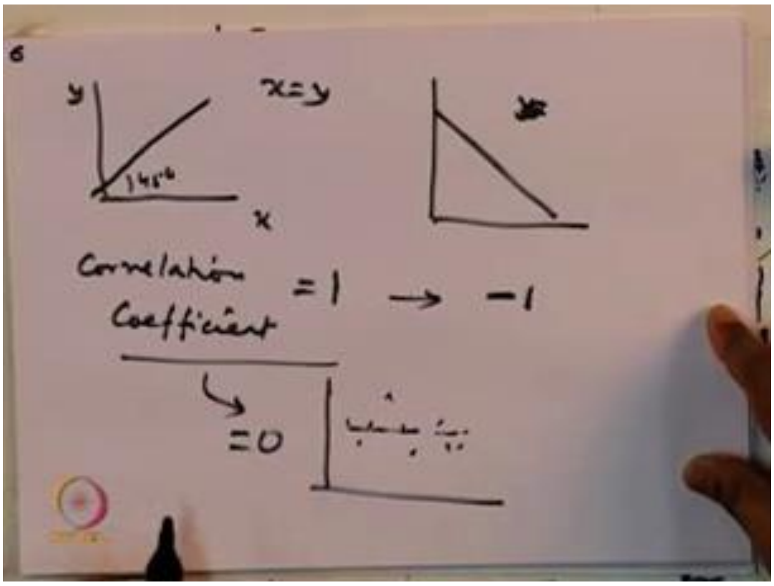


So, you know let us say if I plot my x and y if I plot x and y and I have some scatter plots of some scatter plots like this. So, I can see that on an average if I were to draw a trend line through the middle my trend line will look something like this. So, this is an example of positive correlation.

On the other hand, if my data were to look something like this, this is negatively correlation, this is negative correlation as we saw in the case of weight and fitness, in other cases. So, let us say for example, we are correlating weight with the chance of raining today. So, weight of a person at ten different days and the chance of raining or weight of ten different people and the chance of raining. So, we can clearly see that there is expected to be no correlation between these 2 curves. So, in that case if I draw a line you see that the line will almost look like either horizontal or in some other case it might look almost like this that the line is completely vertical.

So, these are causes where there is no correlation between x and y. So, the mathematical basis for calculating correlation and regression, so, what you have? So, you have this.

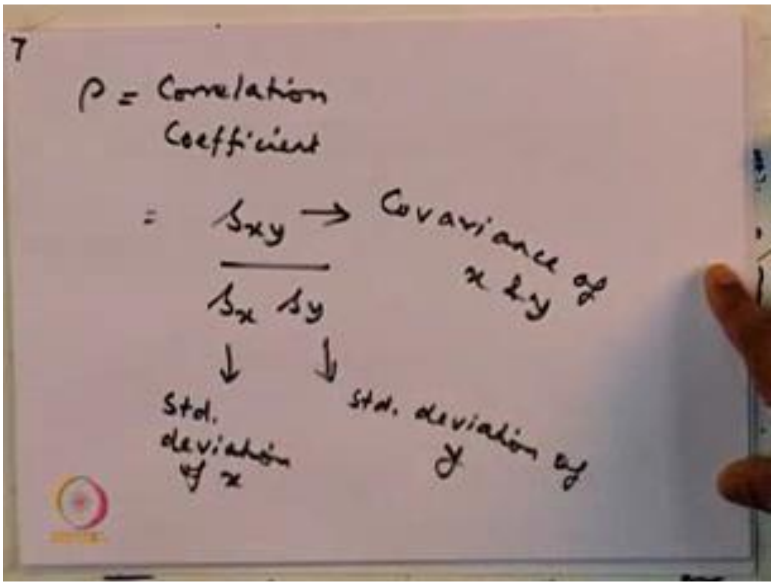
(Refer Slide Time: 09:09)



So, let us just say again in the case let us say x equal to y you have a function which is x equal to y, we know it will be a 45 degree line passing through the origin, this is a case you will come up with something called a correlation coefficient which will come out to be 1.

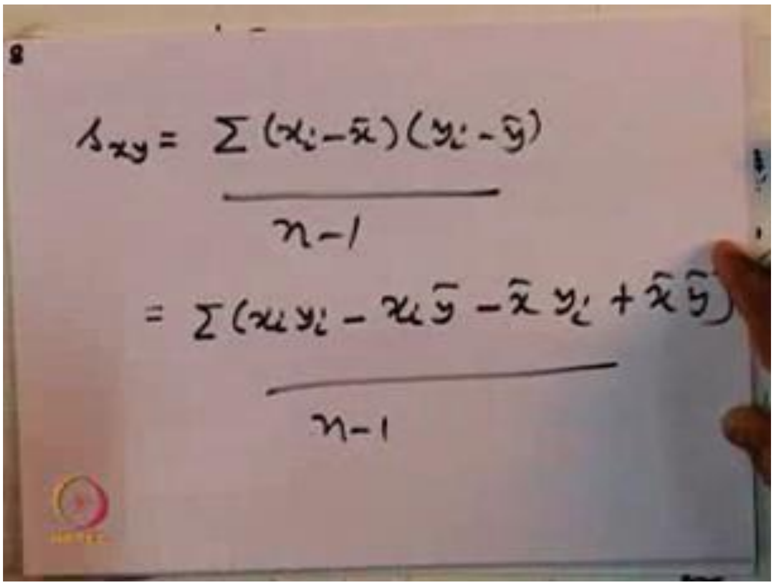
So, in other words, you are they are fully correlated any increase in x will give you the equal increase in y and the other hand, let us say you have a complete opposite slope and this is the case where let us say y is equal to; so in this case, your correlation coefficient is going to be close to value of minus 1 versus when there is no correlation when you have data like this here your correlation coefficient will give a value of 0.

(Refer Slide Time: 10:01)



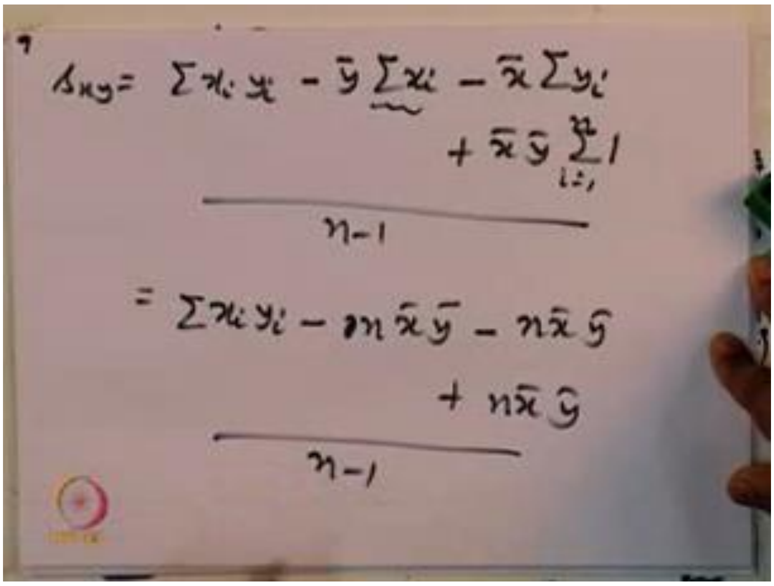
Now, how do you define correlation coefficient mathematically the mathematical definition of correlation coefficient is typically written as rho is represented by rho correlation coefficient is nothing but defined by s x y by s x into s y. So, where s x this is standard deviation of x standard deviation of y and this s x y is called the covariance of x and y.

(Refer Slide Time: 10:48)



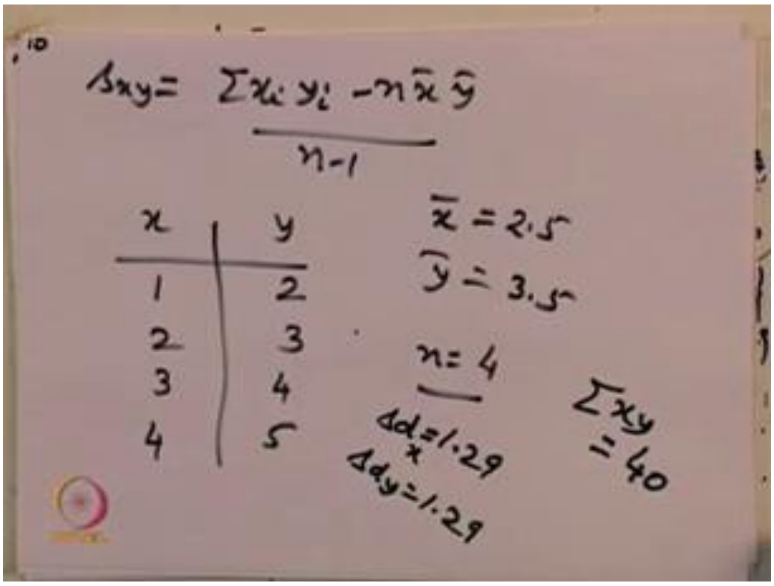
Covariance is defined by s x y is equal to summation of x i minus x bar into y i minus y bar whole divided by n minus 1. So, let us I can expand this further. So, I can expand this to summation of x i y i minus x i y bar minus x bar y i plus x bar y bar by n minus 1.

(Refer Slide Time: 11:25)



So, I know. So, my s x y equal to summation x i y i minus y bar I can take out summation x i minus x bar I can take out summation y i plus x bar y bar summation 1 I equal to 1 to n by n minus 1. So, I can rewrite this as summation x i y i. So, summation x i is nothing but n times x bar. So, I can write this as n x bar y bar minus, similarly here in n x bar y bar plus n x bar y bar by n minus 1.

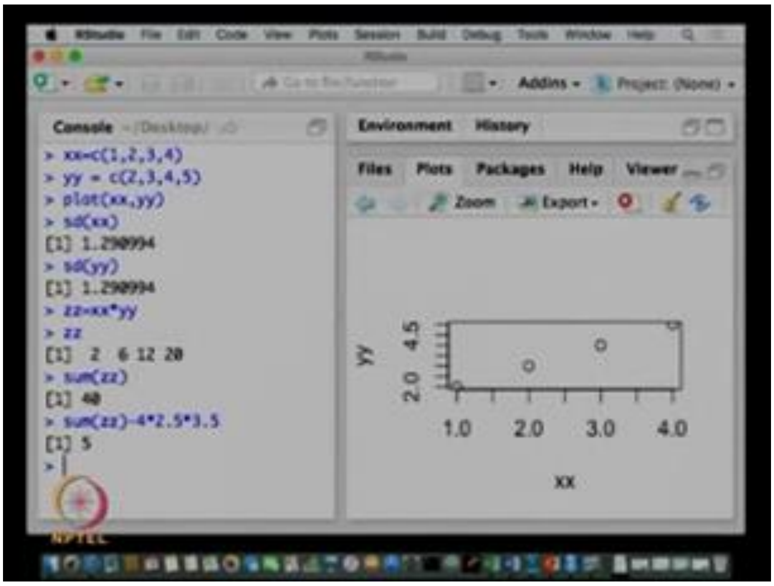
(Refer Slide Time: 12:13)



This gives me to the formula that s x y is summation x i y i minus n x bar y bar by n minus 1. So, this is the difference definition of covariance.

Now, let us generate 2 vectors. So, let us see what kind of covariance we get what is the value of standard deviation and what is the final correlation coefficient for some distributions let us take one particular example where we think that they are positively correlated, let us assume that I have the following 4 values of x and. So, I can calculate what is the value of x bar x bar is equal to 2.5 y bar equal to 3 point 5 I can find out. So, let us open RStudio let me enter.

(Refer Slide Time: 13:22)

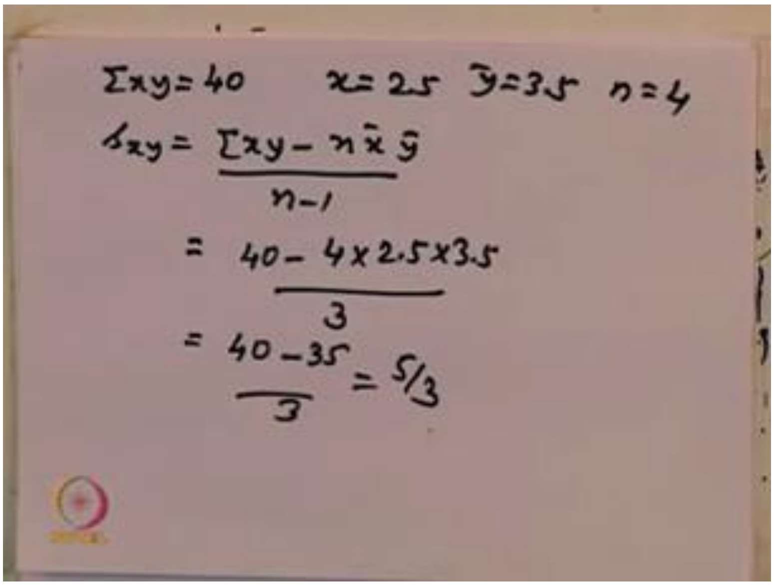


So, let us open RStudio and let me enter x x is equal to sorry c of 1, 2, 3, 4, y y is equal to c of 2, 3, 4, 5, I can plot x x comma y y and this is how my plot looks like, you can clearly see that there is a very linear correlation between x x and y y. So, I want to find out what is the value of s x y. So, I can find out. So, I know that s x y is equal to summation x i y i minus n x bar y bar by n minus 1. So, I have n is equal to 4 in this case.

So, let us calculate the value of s x. So, I can write down here itself I can write down s d of x x s d of y y same thing. So, I can define z z equal to I can define z z equal to let us say s d equal to x x star y y. So, I can find out what is the value of z z which is nothing but summation x y. So, what I have calculated is summation x y then I can add up if I do sum of z z, I get the complete value which is 40. So, I can see that summation x y is coming out to be a value of 40, I know standard deviation of x is equal to standard deviation of x equal to 1.29 and s d of both x and s d of y is equal to 1.29.

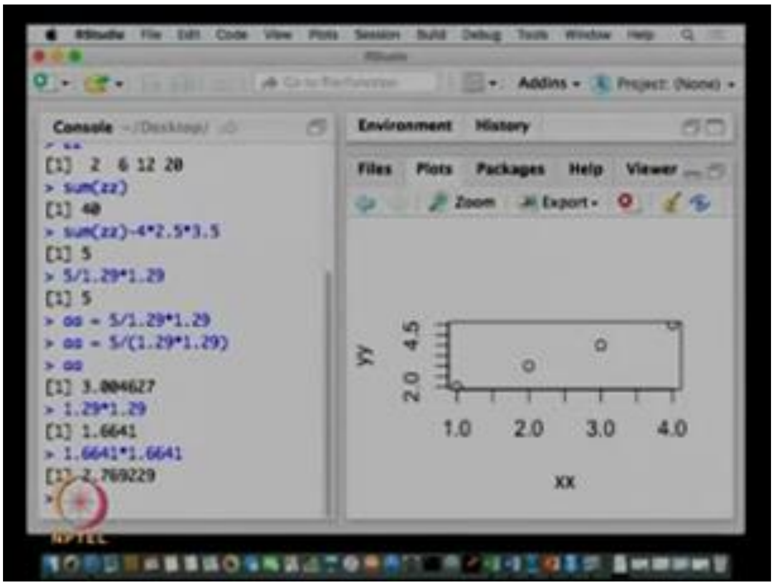
So, now let us calculate. So, I know n is equal to 4. So, sum of zee minus x bar is. So, minus n is 4 star x bar which is 2.5 star y bar which is 3.5 gives me a value of 5. So, I can do s x y. So, I can get the value of s x y.

(Refer Slide Time: 15:45)



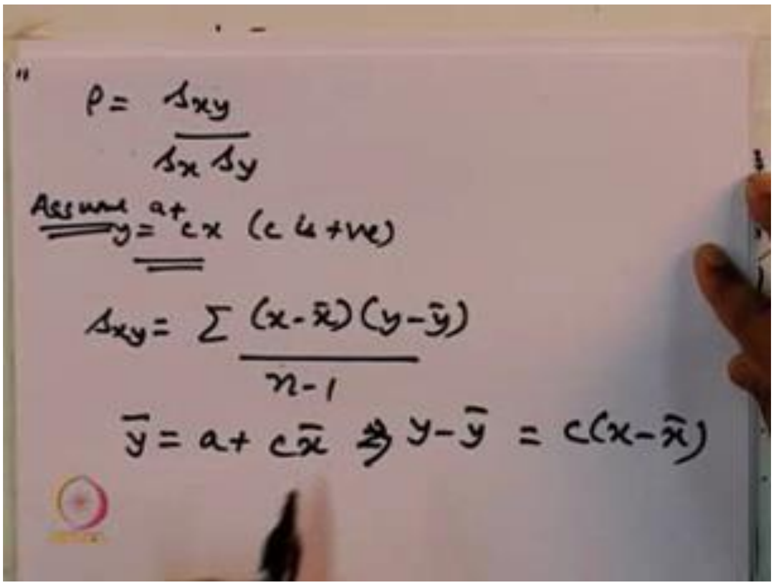
So, I got summation x y is equal to 40 I got. So, s x y is equal to summation x y minus n x bar y bar. So, I have calculated x bar is equal to 2.5 y bar equal to 3.5 n is equal to 4. So, I can calculate the value of s x y is equal to 40 minus 4 into 2.5 into 3.5 by n minus 1 is equal to 3 is equal to 40 minus 35 by 3 equal to 5 by 3.

(Refer Slide Time: 16:27)



And I know what is you know standard deviation of s x. So, if I do s x y; my correlation coefficient is going to be 5 by 1.29 star 1.29. So, you can accordingly use and find out what is the correlation coefficient of rho. So, determine the formula and use to find out the correlation coefficient of y.

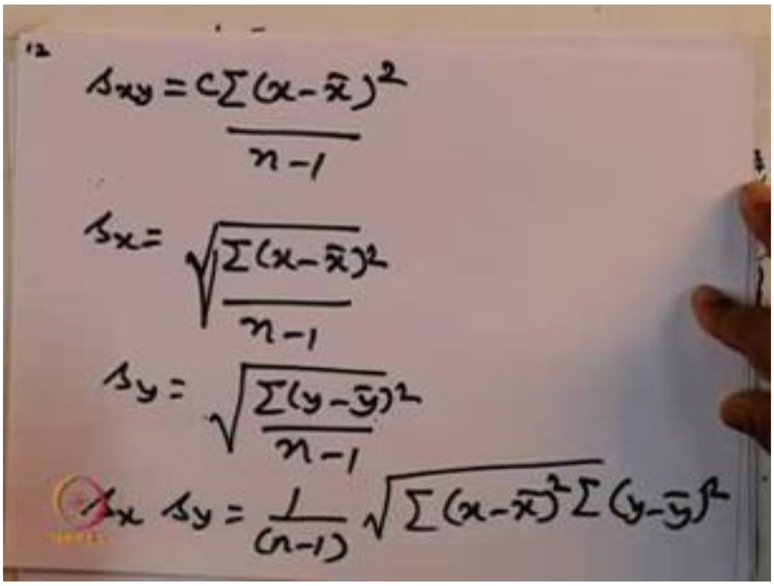
(Refer Slide Time: 17:42)



Now, let us say one thing. So, my correlation coefficient rho is define it by s x y by s x into s y. So, in this case, so my s x y so what is the minimum value of rho possible and what is the maximum value of rho possible. So, we thought that we reasoned that it value this value should be between minus 1 and 1. So, let us see if that is true. So, let us assume y as let us say c x in this case where c is positive let us assume. So, we assume a very strong correlation.

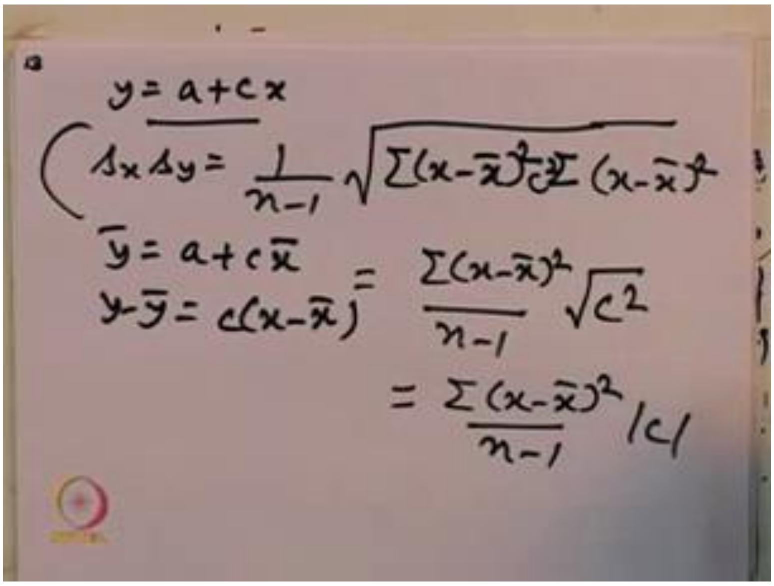
In fact, we can also add some a plus c x to make it more general if we make it assume as y is equal to a plus c x where my s x y is defined as summation of x minus x bar into y minus y bar by n minus 1 right. So, I know my y bar has to be a plus c x bar from previous classes we had derived this equation. So, y minus y bar is nothing but c of x minus x bar. So, this implies y minus by y a c of s x minus x bar.

(Refer Slide Time: 19:03)



So, if this is true, I can compute the value of s x y as summation of x minus x bar. So, I can take c out whole square by n minus 1 c x y is this now s x is root of summation x minus x bar whole square by n minus 1 root of this and s y is equal to root of summation y minus y bar whole square by n minus 1. So, s x into s y will give me. So, I can take out root of n minus 1 I can take out n minus 1 common into root of summation x minus x bar whole square into summation y minus y bar whole square.

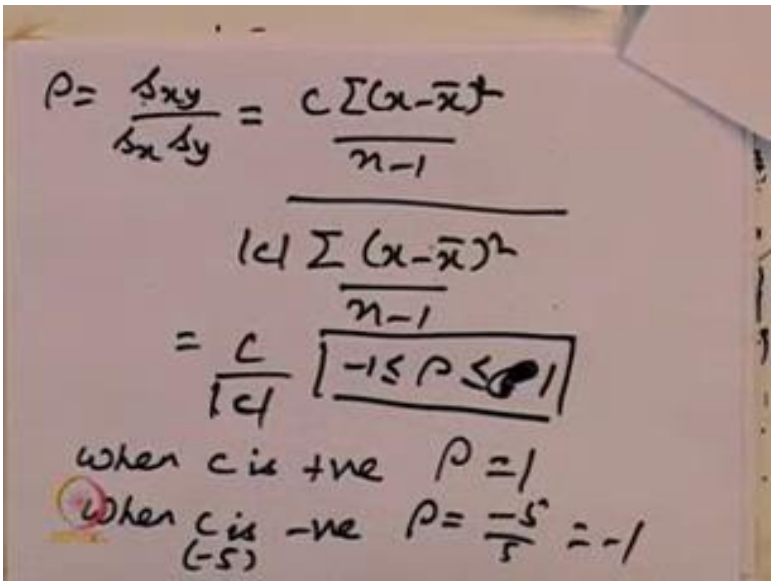
(Refer Slide Time: 19:58)



If this was true then I again know y is equal to a plus c x. So, my s x into x y will be one by n minus 1 into root of summation x minus x bar whole square into summation. So, y is a plus c x. So, again I know y bar is equal to a plus c x bar. So, y minus y bar is equal to c of x minus x bar. So, I can put my c outside. So, c square into x minus x bar whole square.

So, under root I can take it out is equal to summation x minus x bar whole square by n minus 1 into root of c c square. So, this is what I have. So, in the you know I can write it as summation of x minus x bar whole square into n minus 1 into mod of c.

(Refer Slide Time: 21:01)



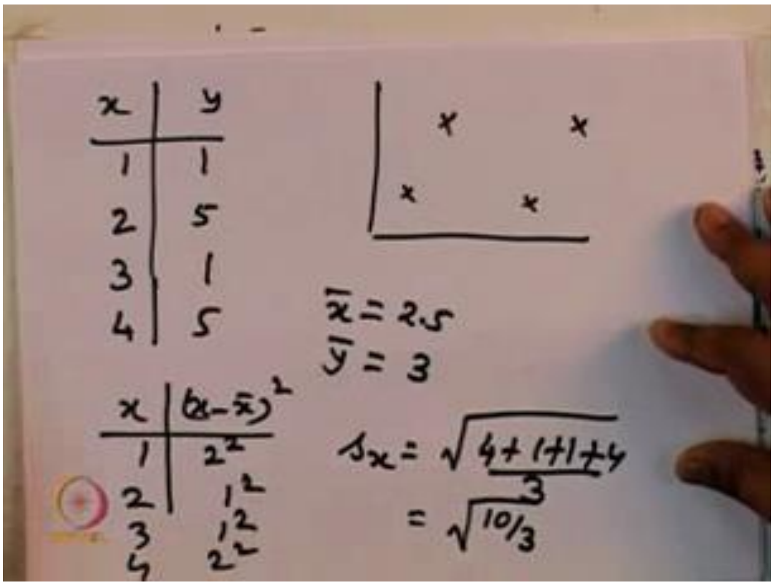
So, if I do this then my s x y. So, my rho is defined as s x y by s x into s y which is going to be is equal to c into summation of x minus x bar whole square by n minus 1 divided by summation. So, mod c summation x minus x bar whole square by n minus 1 is equal to c by mod c.

So, when your c is when c is positive, I can clearly see rho is going to be 1. So, if let us say c is 5 then 5 by mod 5 is simply equal to 1 when c is negative rho is going to be let us say let us say example is minus 5 minus 5 by 5 equal to minus 1 this tells you that your c your rho correlation coefficient is bounded within the following limits rho it bounded between minus 1 and plus 1.

So, this is the value of correlation coefficient. So, minimum correlation coefficient when they are anti correlated; that means, x is increasing y is decreasing or the reverse wave x is decreasing y is increasing when there is complete anti correlation you will get a value of rho which is minus 1 when they are perfectly in sync that is x increases y increases that exact same rate you would get a value of rho is equal to 1. So, when there is some association, but it need not be fully strongly associated you might get a positive correlation or a negative correlation, but the value would be like let say a point 2 or minus point 2 depending on the extent of correlation.

So, now let us say; let us take 1 another example and you know and do this calculation ourselves.

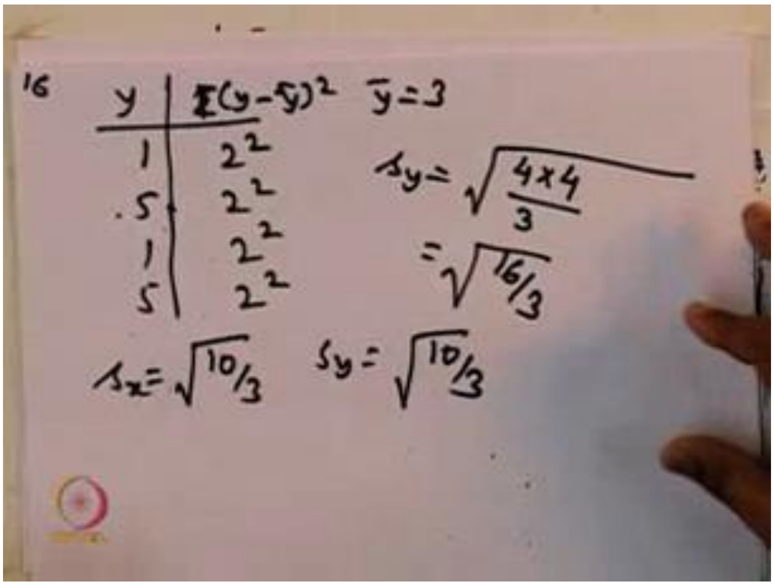
(Refer Slide Time: 23:03)



So, let us take an example where x and y are not particularly correlated let us say if I take this point this point this point and this point. So, let us say y 1 1 x is 2 y is 5 x is 3 y is one x is 4 y is 5 let us do the following exercise.

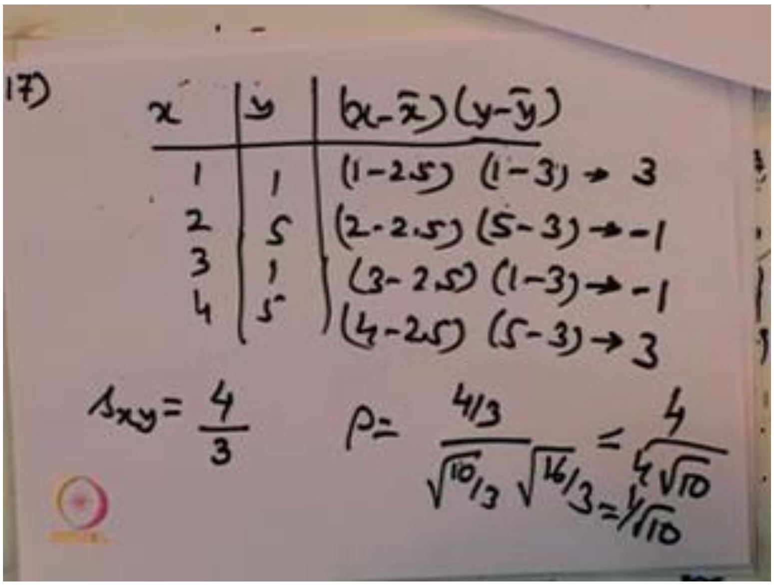
So, my x bar is going to be 2.5 as before y bar is going to be 3. So, let us find out the standard deviation s x. So, if I know the value of x x minus x bar whole square 1, 2, 3, 4. So, this 2 square 1 square 1 square 2 square. So, s x should give me a value of root of 4 plus 1 plus 1 plus 4 by n minus 1 which is 3 is equal to root of 5 plus 5 10 by 3 s x is root of 10 by 3 s x is root of 10 by 3.

(Refer Slide Time: 24:35)



I can do the same thing for y summation y minus sorry y minus y bar whole square I have 1, 5, 1, 5. So, y bar is 3 I know y bar equal to 3 2 square, 2 square, 2 square, 2 square. So, s y should be root of 4 into 4 by 3 solo by 3. So, we have s x is equal to root of 10 by 3 s y is equal to root of 10 by 3.

(Refer Slide Time: 25:24)



Now, we want to find out s x y. So, let us take another piece of page. So, we have x we have y 1, 2, 3, 4, 1, 5, 1, 5 x bar. So, 1 minus 2.5 into 1 minus 3, 2 minus 2.5 into 5 minus 3, 3 minus 2.5 into 1 minus 3, 4 minus 2.5 and 5 minus 3, you can find out this value. So, this is 2 all of them are 2 minus 2 1.5 this is 1.5 to 3 minus into minus is plus this value is doing minus 1 this value is 0.5; 0.53.

So, I can calculate the summation. So, s x y will be summation of this which is 8 minus 2 is 4 by 3. So, my rho is nothing but 4 by 3 by. So, this is 16 by 3 root of 10 by 3 into root of 16 by 3. So, I see. So, my 3 will go is 4 by 4 into root 10 is equal to 1 by root 10. So, I will get a value which is root of 3 is a one-third; roughly one-third. So, you see that this is still positively correlated as per this calculation where it is much lesser than 1, but it is still positive.

So, in this class, we discussed about correlation coefficient and how you can make use of R to calculate these individual metrics and even calculate the correlation coefficient. In the next class, we will again take few more examples of correlation and then go to the next step of how to do regression and fitting.

With that I end here and I look forward to meeting you in the next lecture.

Thank you.