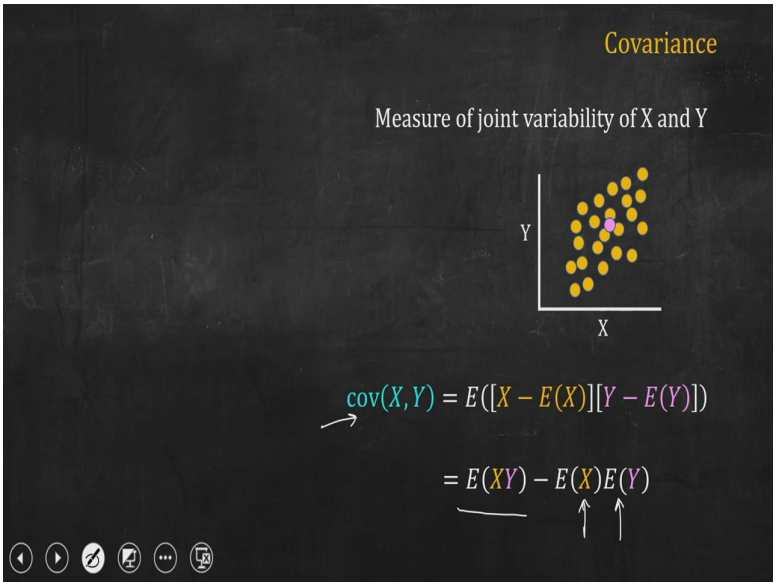
Data Analysis for Biologists
Professor Biplab Bose
Department of Biosciences & Bioengineering
Mehta Family School of Data Science & Artificial Intelligence
Indian Institute of Technology, Guwahati
Lecture – 26
Correlations

(Refer Slide Time: 00:48)

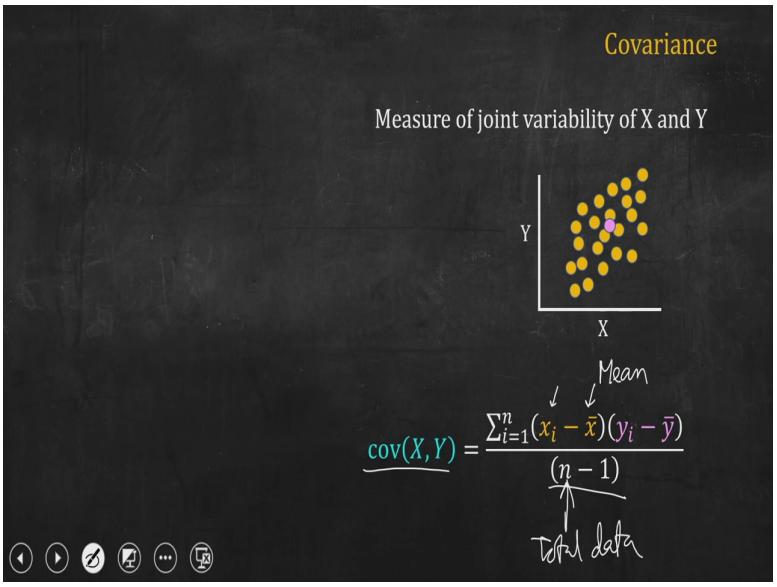


$$cov(X, Y) = E([X - E(X)][Y - E(Y)]) = E(XY) - E(X)E(Y)$$

Hello everyone. In this lecture, we will learn about Correlation, Correlation between variables in our data set. But let me start the lecture with covariance. In an earlier lecture, we have learnt about covariance. So, if I have two variable X and Y, you have measured them in an experiment and you may have represented the data in a scatter plot or something like that and you want to measure the covariance of that.

By definition of covariance, the covariance between X and Y as we know by definition from the probability theory, we get it is equal to expectation of X into Y, the multiplication of X into Y and the expectation of that minus expectation of X into expectation of Y. Their expectation is actually the arithmetic mean.

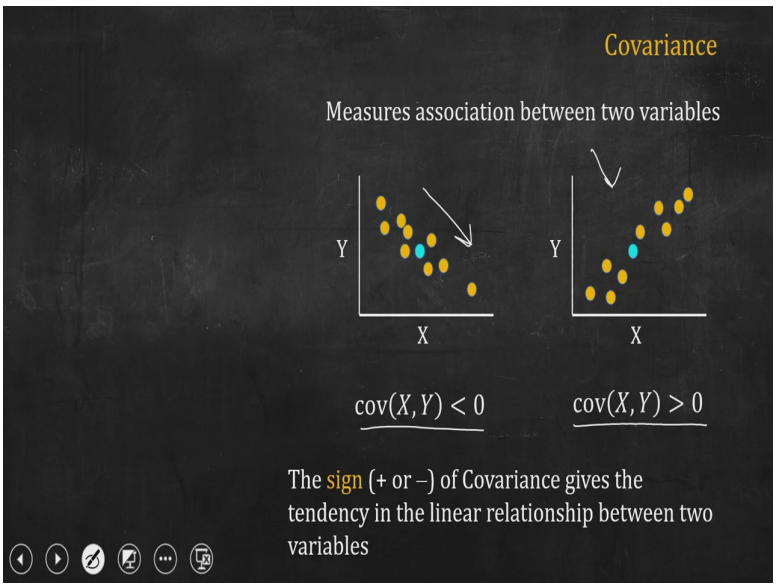
(Refer Slide Time: 1:34)



$$cov(X, Y) = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{(n-1)}$$

Now if you have a sample data, the handful data points, we do not use this rule rather we use a formula derived from this rule and corrected for small sample size and we have learnt that earlier also and in that case the covariance or sample covariance the way we see say, is equal to what you do, you take the particular measurement of x subtract from that the mean of x that is the mean, x bar into y minus y bar and you sum all this multiplication together for all the data point and then divide it by n minus 1 where n is the total number of data, total data points. That is what we have learned and when you are talking about, in that lecture we were talking about covariance. We also learned that covariance is a measure of a joint variation between two random variable that is good enough, fair enough.

(Refer Slide Time: 2:25)



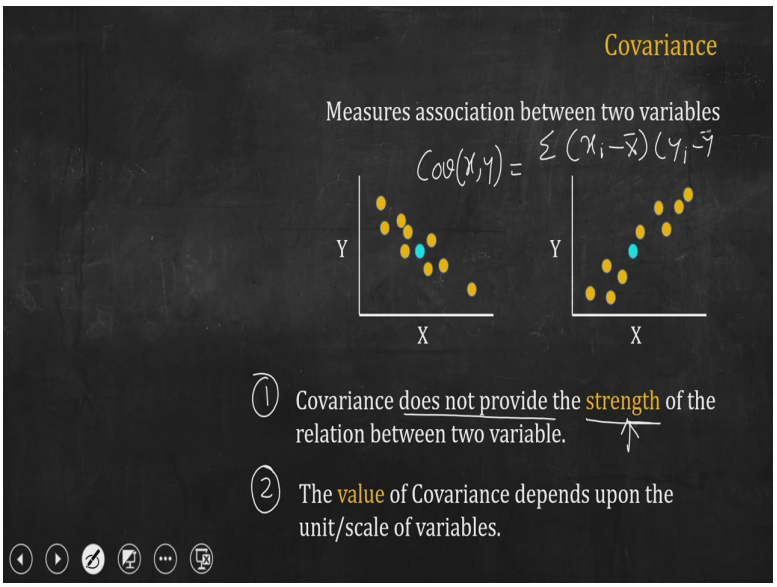
$$Plot\ 1:\ cov(X,\ Y) <\ 0$$

$$Plot\ 2:\ cov(X,\ Y) >\ 0$$

But it also tells us something more, it tells us something about the linear relationship between two variable this x and y in this case. So, what do I mean by that? When we have a negative linear relation that means when x increases y decreases. If I have that situation sort of linear trend is there decreasing trend is there. In that case it can be shown that covariance will be negative. Whereas, if I have a trend positive trend means when X increases Y also increases some sort of linearly, then I will have this situation here, in this second plot and you will find the covariance between x and y will be bigger than 0. So, it is positive.

So, that is what we said in that lecture that the sign, the sign of the covariance gives us the linear trend of association between two variables. Now, this is a good indication of some sort of linear association between two variables in my data set. But this covariance is not really the correct or perfect measure for linear association between two variables. Because there are two reason let me explain why it is not so useful for us, although the sign tells me about some sort of linear trend positive or negative.

(Refer Slide Time: 3:42)



$$cov(X, Y) = \sum\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)$$

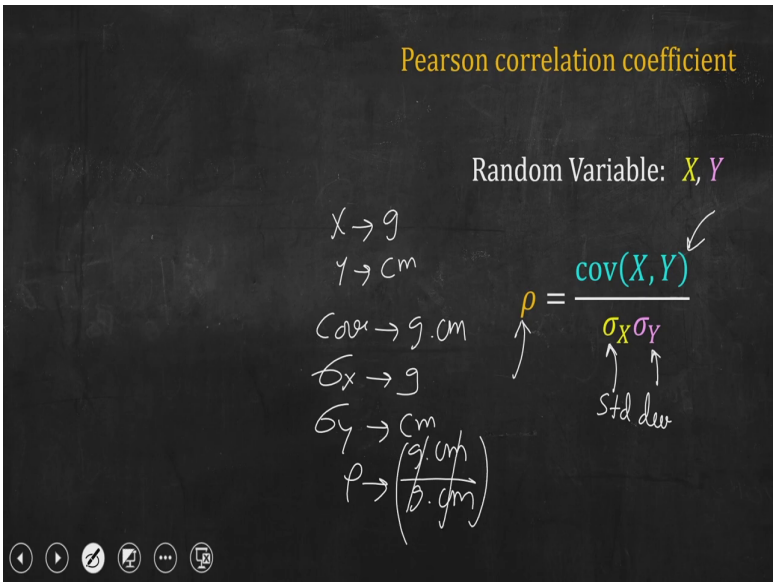
The first problem is, it tells me the trend but it does not tells me the strength of that linear association. So, the sign tells me whether it is a positive linear relation or a negative linear relation but it does not tell the strength of that linear association. So, the first problem in using covariance to identify association between any two variables in your data set or two variables that you may have measured is that it does not provide the strength of that linear association, that is one problem. It does not tells me the strength of that association.

Second thing is that, remember the covariance depends upon the unit or scale of your measurement. Let me explain it bit, let me explain it here itself. So, if I write covariance between x and y, I will use that formula that we have we used for a handful of data set in a sample. So, we do summation into x i minus x bar into y i minus y bar and then we divide it by n minus 1. I am just writing here, the denominator is not so important here, in this particular issue.

See, suppose x is measured in gram and y is measured in meter or centimetre something like that, so x is weight, y is length. So, this one has unit of gram, whereas this one has unit of suppose centimetre. So, what is the unit of covariance? Gram into centimetre. Now, suppose you have another third, or other experiment where you have measure to taken two variables which are both of them are length. So, then you have centimetre square for covariance. So, that means, the unit or scale that you are using affects the measurement of covariance because if I change gram to pound or kilogram or milligram my covariance value will change.

So, the covariance, the value of covariance depends upon unit or scale of the variables that we are dealing with. But we do not want it because if the measurement of strength of association keeps on changing depending upon which scale or which unit I am using, then I am in trouble. I will not be able to compare with the different data sets. So, I want to get rid of these two problem and Pearson's correlation coefficient that we will learn in this lecture, helps us in solving these two issues.

(Refer Slide Time: 6:31)



$$\rho = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$$

So, let me explain first what Pearson correlation coefficient is. Pearson correlation coefficient takes two component. One it takes covariance at the same time it takes the standard deviation of each of the variables. So, Pearson correlation coefficient often written as rho or small r sometime capital R is also written. So, if some people want that usually we call it rho, so the numerator is covariance between x and y and then we divide that by standard deviation of x and standard deviation of y.

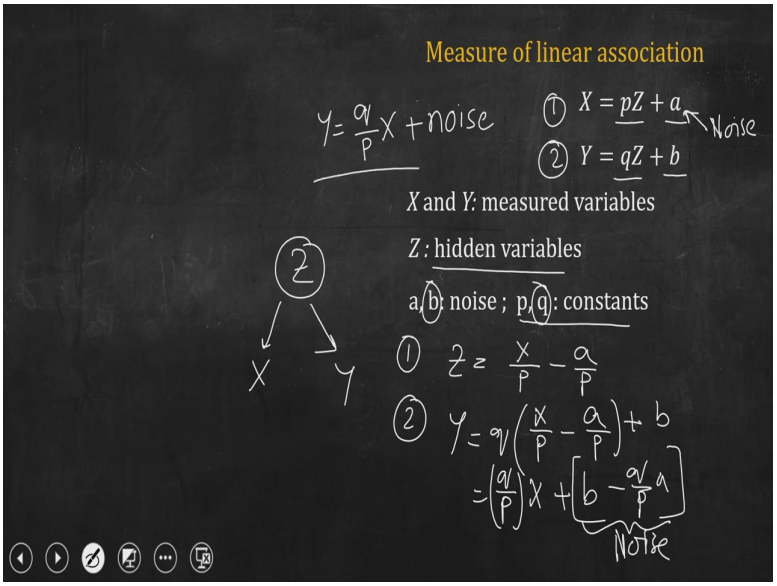
So, it is a normalized thing, you are normalizing the covariance between these two variable x and y by their standard deviations product and this normalization immediately solves the second problem associated with covariance that the dependence on unit, rho the Pearson correlation coefficient is not dependent on unit. Let us see why?

Suppose x unit is, the unit of x is gram and your y units is in centimetre. So, then what is happening here is that covariance will have unit of gram and centimetre whereas your sigma x, the standard deviation of x will have gram will be unit. Standard deviation of y will have

centimetre as unit. So, what is the unit of rho? Unit of rho will be gram centimetre divided by gram centimetre. So, everything get cancelled, I have a unit less term.

So, the second problem that I mentioned for covariance is not there in Pearson correlation coefficient. So, it is unit less, so I can compare different data sets different variables rho without any hesitation. So, it solves the second problem.

(Refer Slide Time: 8:46)



$$y = \frac{q}{p}X + noise$$

$$1.\ X = pZ + a \quad 2.\ Y = qZ + b$$

$$1.\ Z = \frac{X}{p} - \frac{a}{p}$$

$$2.\ Y = \left(\frac{q}{p}\right)X + \left[b - \frac{q}{p}a\right]$$

Now, let me go back to the first problem. How does the Pearson correlation coefficient tells me the strength of linear association between two variable x and y. I will explain that now. So, let us first try to understand what do I mean by linear association between two variables. So, suppose I am measuring x and y, two variables in an experiment. These two variables can be some genes or maybe some other physiological parameter of a patient or something else.

So, x and y are two variable that we are measuring. Now suppose there is a hidden variable Z, that is what I have written Z is a hidden variable. You do not see that variable or you do not measure that variable but Z affects or controls both x and y. So, this is the situation that I have and how Z control x and y that control is linear and that is what I have written here in the first equation and in the second equation.

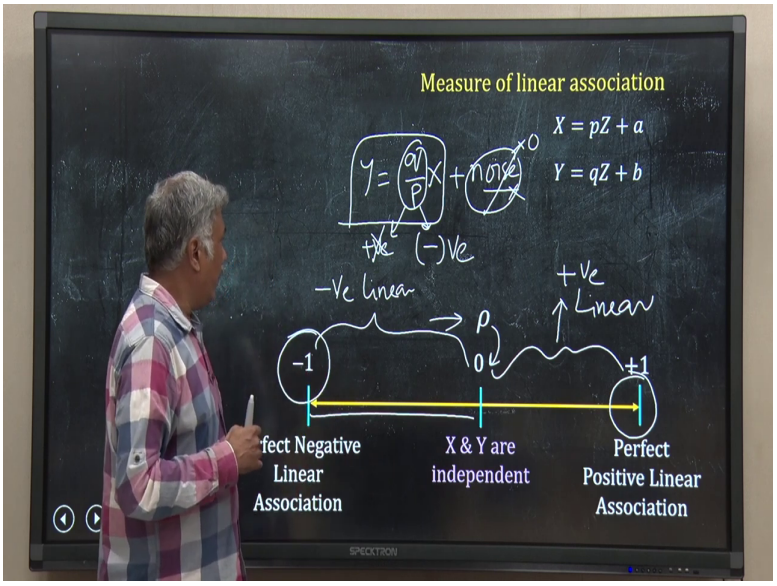
In the first equation it shows how Z controls the value of x. So, x is equal to p into Z where p is a constant plus a, a is a noise. Remember every time you do an experiment, every time you take some measurement in your machine any experiment you do there will be some noise. So, we are considering a as noise. So, the first equation represent how linearly Z controls or affects x. Second one is for y and Z relation. Again, linear so y is equal to q into Z plus b. Again, q is the constant and b is a noise term I have added.

So, now let me arrange this equation 1 and 2 and directly write y in terms of x. So, from the first equation what I will get? I will get Z equal to, if I rearrange term, I will get Z equal to x by p minus a by p. Hope I am right? So, now I will put this value of Z into the second equation. I will replace Z in the second equation.

So, in the second equation what I will get? I will get y equal to q into Z is replaced by x by p minus a by p plus b. So, then what do I get? I get q by p into x plus b minus q by p into a. This one, so the second term this one is nothing but noise because it has a and b. So, I can simply actually consider that as a noise term, we can replace it by any other term. So, what we have got? If I write it clearly here, I have got y equal to q by p into x plus noise.

So, that means you can easily see here I have a linear relation between x and y because Z linearly control both x and y. Now, let me go into how Pearson correlation coefficient captured this linear relation. I will not go into details of the mathematical derivation but from this the part that I have done till now, from this if you continue, use the definition of Pearson correlation coefficient in terms of covariance and standard deviation you will land up what I will reach and what I will discuss now.

(Refer Slide Time: 12:35)



$$1.\ X = pZ + a \quad 2.\ Y = qZ + b$$

$$y = \frac{q}{p}X + noise$$

So, it can be shown that the Pearson correlation coefficient will vary between minus 1 to plus 1 and when I will get plus 1? Let me rewrite what I have got in the last slide. I have got y equal to q by p into x plus noise term. When I have no noise in the system, imagine that is quite imaginary because I keep on saying that in all your experiment will have noise but imagine you do not have noise. Perfect data you have, so in that case this will get cancelled. So, that will be 0 so then you have y equal to q by p into x.

Now, if q by p, if this q by p is positive and your noise is 0, then I will have a perfect positive linear relation and rho the Pearson correlation coefficient will give you a plus 1 value. Whereas, if it is not positive relation, if it is a negative relation means q by p is negative and there is no noise, perfect linear association but negative linear association. Then Pearson correlation coefficient will be minus 1 and when x and y are independent that means we do not hold this relation at all, x and y are not connected, they are independent then I will get Pearson correlation coefficient as 0.

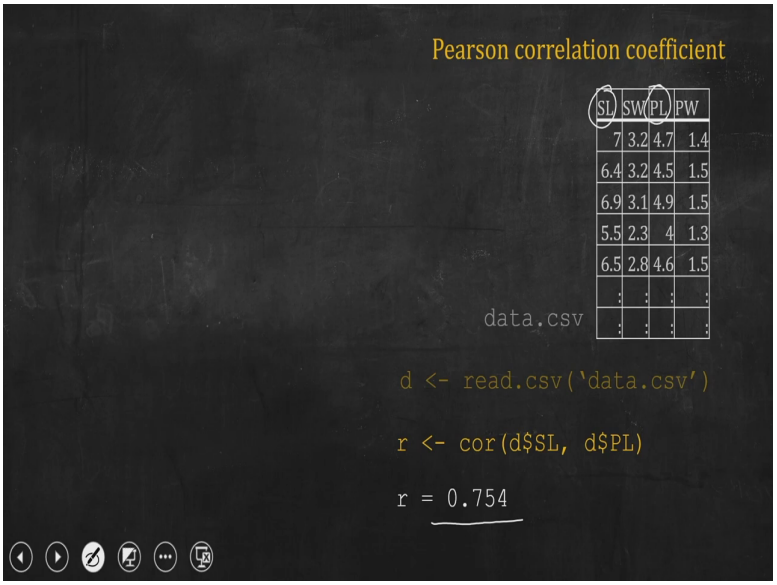
So, I have talked about both the extremum and the middle. When there is perfect negative linear relation between x and y in absence of noise, I get minus 1, when perfect positive linear relation between x and y in absence of noise, I get plus 1 as a value of Pearson correlation coefficient, the rho when x and y are independent. I am in the middle at 0, but as I said that most of our experiments will have some amount of noise that means our most of the

data if I have linear relation, positive linear relation between x and y will be in this region because there will be noise, so they will get shifted from plus 1.

Whereas if I have noise in the system which will happen in all the real cases and there is a negative relation, linear relation between x and y then I will be between minus 1 and 0, somewhere between minus 1 and 0. So, if my Pearson correlation coefficient value rho is close to plus 1, then I will say, I have a strong positive correlation or strong positive linear association between x and y.

On the other hand, if the rho the Pearson correlation coefficient is close to minus 1, then I will conclude that my two variable x and y has negative linear association. So, Pearson correlation coefficient, the value of that where it is in this scale from minus 1 to plus 1, tells me about the strength of linear association. Obviously, also the direction whether it is positive or negative.

(Refer Slide Time: 16:06)



d ← read.csv("data.csv")

r ← cor(dSL, dPL)

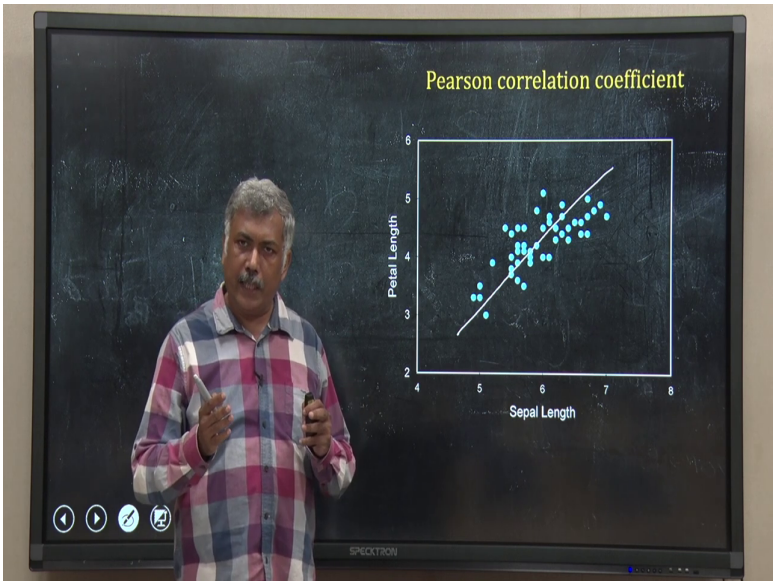
Now, how should I do this? All these things are actually discussion about what is the definition and how we explain and how we use it. So, in reality almost all statistical software will have tool to calculate correlation coefficient, Pearson correlation coefficient in r as we will be we are using r in our course, in r you have to use the correlation function cor function.

So, suppose if I take a data set, let me take the data set from the iris data set, a snapshot of that I have shown here. So, you have sepal length, petal length, petal width are the columns, different variables. And first obviously, I will read that data. So, now this small d is my data set and I want to find the Pearson correlation between two variable suppose.

So, I want to calculate the Pearson correlation between sepal length and petal length suppose. So, how should I use the cor function? I will use this way, I will write cor in that I will use the sepal length variable as one argument and the petal length variable as the other argument and r will collect that value, the correlation coefficient will be collected there I am representing rho by r in this case.

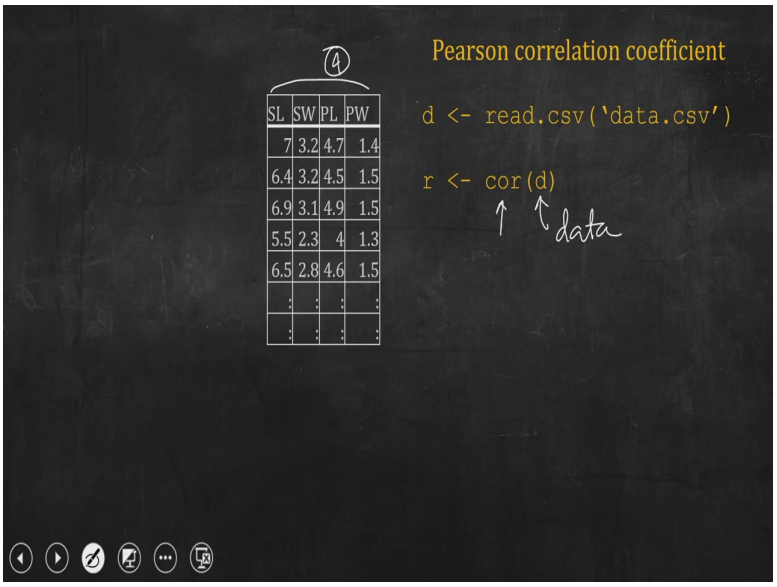
So, I have done that for this particular data set, although I have not shown the data, the complete data set here and it turns out to be the Pearson correlation coefficient comes out to be 0.754, 0.754 is reasonably close to plus 1. So, that means the rho, the Pearson correlation coefficient in this case is saying that there is some sort of positive linear association between sepal length and petal length in this data set, particular data set.

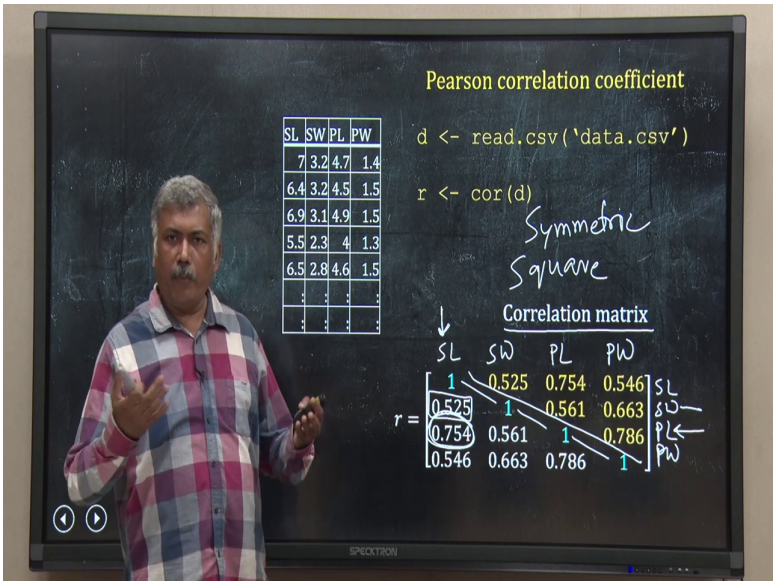
(Refer Slide Time: 17:54)



So, let me plot this, I have made a rough plot for that and if you look into the data, you can easily see there is a rough positive linear relation between the sepal length and petal length and that is why that has got captured by my calculation for rho which has come out to 0.754. Now, the way I have used the cor function, I am calculating the rho between two variables. But suppose I want to do that for all pair of variables in my data set in one single go, how should I do that and the same cor function will actually allow me to do that.

(Refer Slide Time: 18:30)





SL	SW	PL	PW
7	3.2	4.7	1.4
6.4	3.2	4.5	1.5
6.9	3.1	4.9	1.5
5.5	2.3	4	1.3
6.5	2.8	4.6	1.5
:	:	:	:
:	:	:	:

d ← read.csv(“data.csv”)

r ← cor(dSL, dPL)

r =

SL	SW	PL	PW

1	0.525	0.754	0.546	**SL**
0.525	1	0.561	0.663	**SW**
0.754	0.561	1	0.786	**PL**
0.546	0.663	0.786	1	**PW**

So, I have four variables here, four variables. So, I want to calculate the pairwise correlation between all of these, all the pair between of these four variables. So, I will again use the cor function but in this case I will not declare variable separately but rather I will give the whole data here. So, I will write cor d, d as the argument the data as an argument.

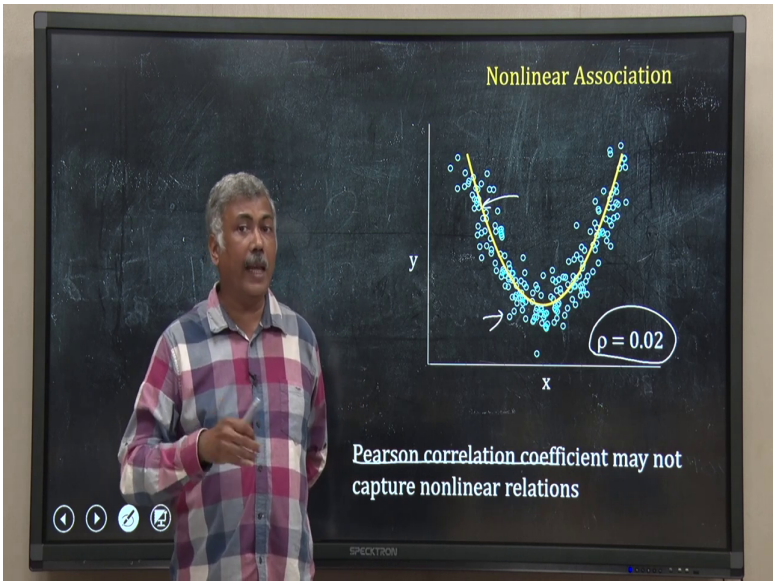
So, the output of this instruction will be actually a matrix and this matrix is called correlation matrix. What is there in the correlation matrix? Correlation matrix is actually giving me correlation between any two pair of the variables in my system. For example, in this case I have SL here, sepal width here, petal length here, petal width here. This one is also the same sequence, sepal length, sepal width, petal length, petal width. So, what is this 0.525? 0.525 is the Pearson correlation coefficient between sepal length and sepal width.

Now, just few slide back, I have calculated the correlation coefficient between sepal length and petal length and it turned out to be something like 0.754. So, let us check sepal length and petal length. So, I get this one, so this is the value of correlation between SL and PL. Pay attention to this diagonal. These are always be 1, because the correlation between same variable x and x, y and y, Z and Z will be always 1.

So, in a correlation matrix the diagonal element will be always 1. Pay attention to two other thing also, correlation matrix is a square matrix. So, this is a square matrix, at the same time it is also a symmetric matrix that is why I have shown this part in yellow, one part in yellow, the part above the diagonal I have shown in yellow.

You can see I have the mirror image on the other side. So, if I take a transpose of the matrix I will get the same thing. So, correlation matrix is actually a symmetric square matrix and it has all the properties of symmetric square matrix, those are very useful in our data analysis algorithms. So, this is how in one single line statement, I can get all the correlation between all pairs of variables in my data set. So, that is quite handy actually.

(Refer Slide Time: 21:33)

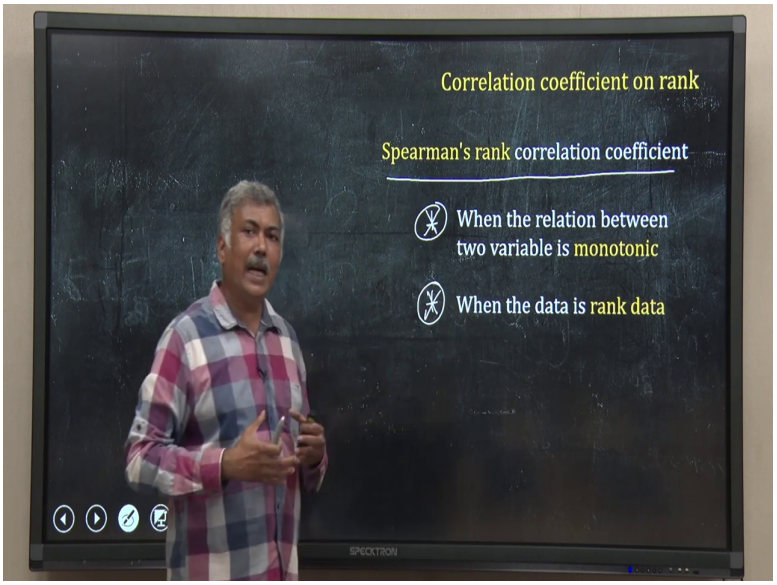


Now, till now what I have discussed is about only Pearson correlation coefficient and every time I am talking about Pearson correlation coefficient, I am saying that it measures a linear association between two variables. It is 0 only when x and y are independent. Actually, Pearson correlation coefficient has trouble with nonlinear data because if you remember when we were discussing about covariance, we have shown one example where I have a non-linear data and that has covariance of 0. So, that means some non-linear data set can have confounding covariance and that will lead to confounding confusing Pearson correlation coefficient also.

For example, what I have shown here, I have a non-linear data set, x and y. These blue dots are each data point. And I have fitted a or rather I overlaid a yellow guide line which as a whole shows the trend in the data and you can easily see there is a non-linear smooth non-linear trend actually. Even without this yellow line you can understand that.

Now, if you measure the Pearson correlation coefficient for this data, it will be 0.02 and if you have not seen the data, if you have not seen that only the table is given to you and it just calculated using r, the correlation coefficient between x and y you will say there is no relation, there is no association between x and y. But there is an association hidden here and once you plot it you immediately recognize there is a non-linear association between x and y and your Pearson correlation coefficient is failing to capture that. So, how to solve this problem?

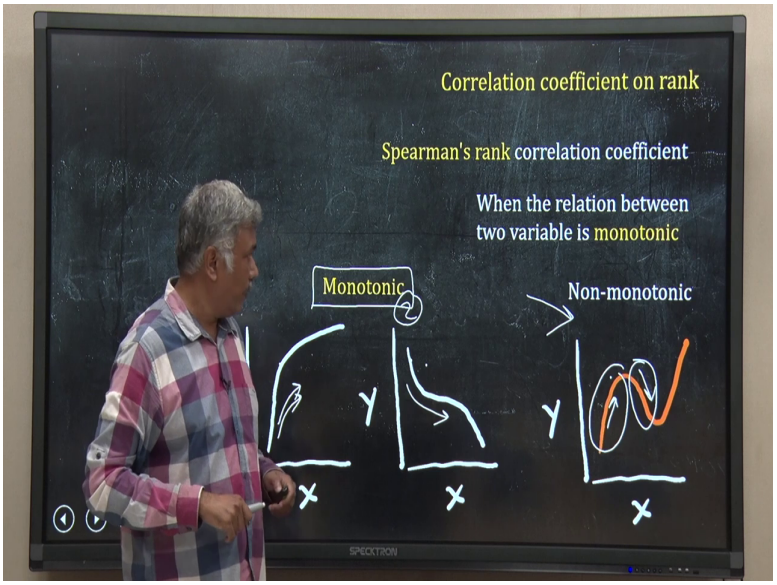
(Refer Slide Time: 23:28)



This problem of non-linear data can be sorted out to some extent, if I use another type of correlation coefficient called Spearman correlation coefficient. It is actually called many times Spearman's rank correlation coefficient and it has primarily two purpose. The first is, when I have some sort of monotonic relation between variables which may be linear or non-linear does not matter, we should use Spearman's rank correlation coefficient.

The second utility of this correlation coefficient is that, suppose the data is itself a rank data, mean category data. At least one variable is rank data suppose, for example you are measuring something as bad, good, ugly. So, three categories, so you may give mark them as 1, 2, 3, scores. So, if you have that type of data, you cannot use Pearson correlation coefficient. You have to use something which is called rank correlation coefficient and Spearman's rank correlation coefficient is one of them.

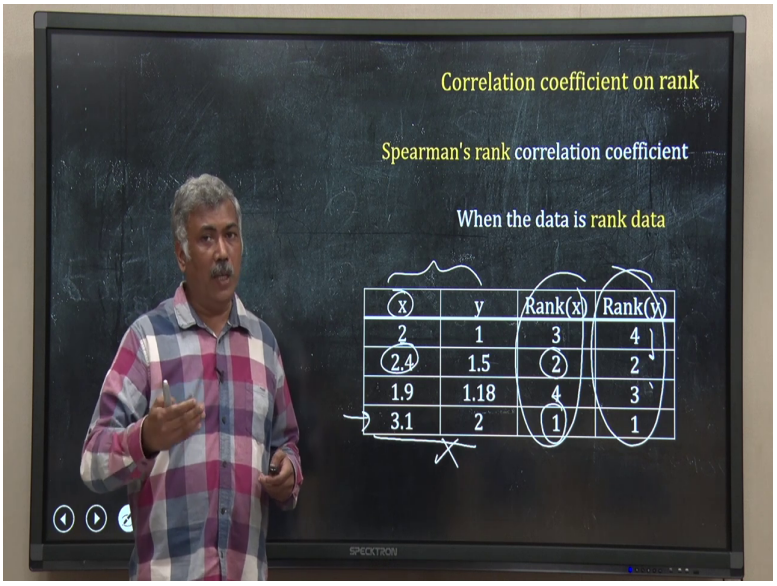
(Refer Slide Time: 24:29)



So, let me first explain what do I mean by monotonic relationship. I have drawn a multiple figure here, take the first figure. In this case, as x increases y also increases, the slope may change. How fast y is increasing with respect to x may change but always monotonically as x progresses become higher, y also increases. Whereas look at this one, this is called non-monotonic because in this region as x increases y increases but after sometime the slope changes the sign and as x increases y decreases and then again it swipes back.

So, this is non-monotonic. Let me give another example of monotonic one, this second drawing here. In this case, as x decreases y also decreases. The slope is not same always. If the slope is constant then I have got a straight line. Slope is not constant but there is a monotonic change of y with respect to x. So, Spearman's rank correlation coefficient are good to detect this type of monotonic association between two variables.

(Refer Slide Time: 25:44)



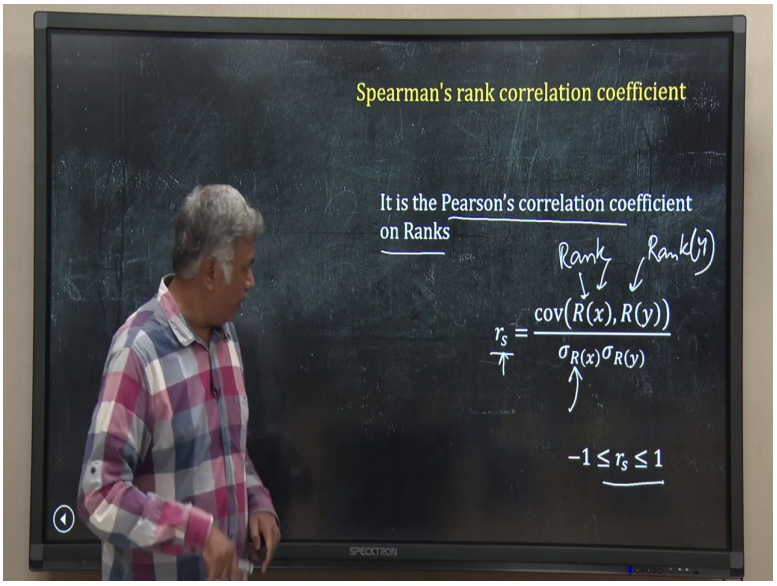
x	y	Rank(x)	Rank(y)
2	1	3	4
2.4	1.5	2	2
1.9	1.18	4	3
3.1	2	1	1

Let us look at the second issue of rank data, what do I mean by rank data? Suppose, this is my original raw data, x and y two variables I have measured. Now, looking to the data of x, I have 2, 2.4, 1.9 and 3.1. Who is the largest one? 3.1 is the largest one. So, I say its rank is 1. Which is the second largest number? 2.4, so, its rank is 2. In this way, I give rank to each of the values of x. So, I get 3, 2, 4, 1.

Similarly, I rank the data of y original data of y. So, I get 4, 2, 3, 1. In this case, the highest value I have considered as rank 1. You may consider the lowest value as rank 1 also, it does not matter. So, what we have done? We have arranged this data as per their value, ascending or descending order. And we have a rank.

Now, I want to calculate the correlation between this rank data, not the original data. And Spearman correlation coefficient, rank correlation coefficient actually use this rank data either you have the rank data from the very beginning or you convert raw data into ranked data, the way I have shown and then you use Spearman correlation coefficient on that.

(Refer Slide Time: 27:10)



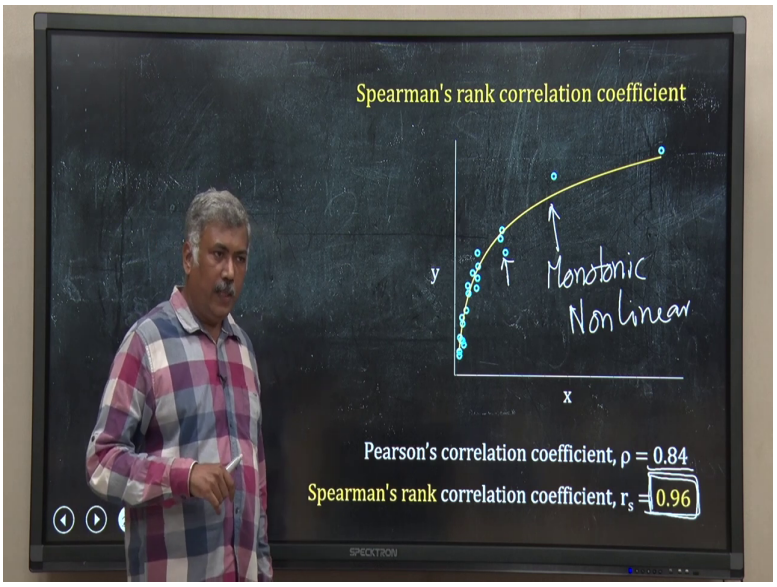
$$r_s = \frac{cov(R(x), R(y))}{\sigma_{R(x)}\sigma_{R(y)}}$$

$$, -1 \leq r_s \leq 1$$

So, what is the definition of that? It is very simple, Spearman's rank correlation coefficient is nothing but a Pearson correlation coefficient but on ranks, not on raw normal data that we use. So, what we will use? In the numerator we will have, usually we write rs to represent this as the Spearman's rank correlation coefficient. We will have covariance between not between x and y but between rank of x, so ranked data and rank y.

So, we will take the ranked data and calculate the covariance between those two and will divide by not the standard deviation of x and y rather the standard deviation of rank of x and rank of y. So, it is exactly same as Pearson correlation coefficient but we are using the ranked data and it can be shown that this rs will vary from plus 1 to minus 1 and it has the similar meaning that if there is a monotonic positive relation between two variable x and y, it will be close to 1. If it has a negative monotonic relation between x and y it will be close to minus 1.

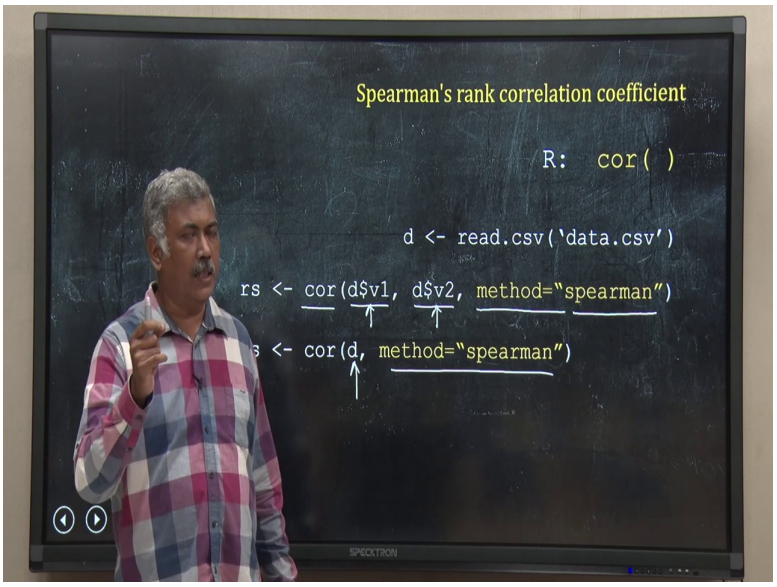
(Refer Slide Time: 28:24)



Now let me show one data for that. So, the blue line dots, blue dots are again the data point and I have overlaid a smooth non-linear line as a guideline to highlight how the data varies and you can easily recognize that this is actually a monotonic non-linear relation. Monotonic non-linear relation, it is non-linear. I have used Pearson correlation coefficient on this, calculated that and it turns out to be 0.84.

Now, if I calculate Spearman rank correlation coefficient rs, it turns out to be 0.96. So, the correlation coefficient has improved. It has improved because the data is actually non-linear, so Pearson correlation coefficient does not capture, does not do justice to the correlation in that data. So, Spearman correlation coefficient is better to use in this case and it is suitable because the data is non-linear but monotonic and that is why we have a better value of 0.96 and easily looking at the data even without plotting it. You can say, I have a good correlation between x and y.

(Refer Slide Time: 29:56)



d ← read.csv("data.csv")

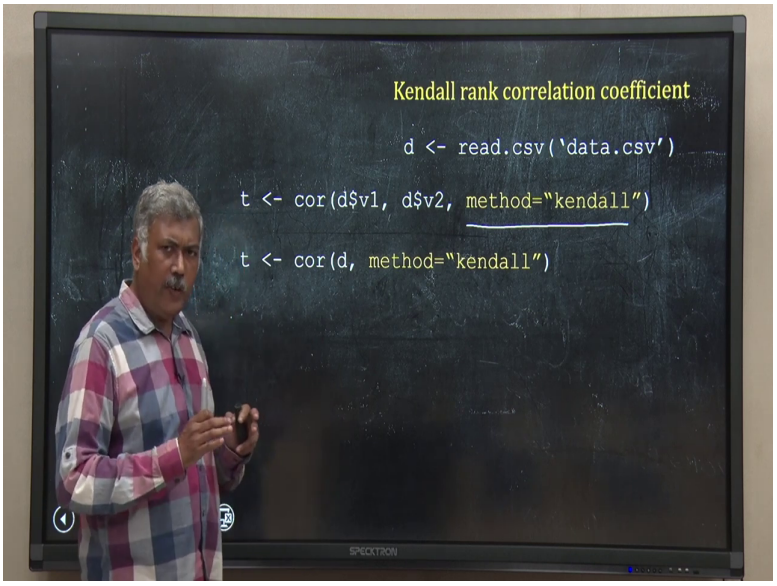
rs ← cor(d$v1, d$v2, method = "spearman")

rs ← cor(d, method = "spearman")

Now, how do I calculate this? Thankfully the cor function of r by can calculate very easily the Spearman correlation also. Actually, the cor function I can give argument methods of correlation calculation as arguments. By default, it uses the Pearson correlation coefficient as a method but I can define the method to be used. For example, for the same data set, I can write cor and then I give the first variable and give the second variable as argument and then I can write method equal to Spearman.

Then the cor function will use Spearman method, definition of correlation function and it will calculate the Spearman's rank correlation coefficient for these two variable v1 and v2. Similarly, if I want to calculate the correlation matrix of the whole data set but I want to use Spearman's method or Spearman's correlation coefficient. Then I will again say method equal to Spearman and I will take the whole data set as the argument. So, I will get the correlation matrix. Interestingly, cor function has another method, the first method is obviously by default is Pearson.

(Refer Slide Time: 31:07)



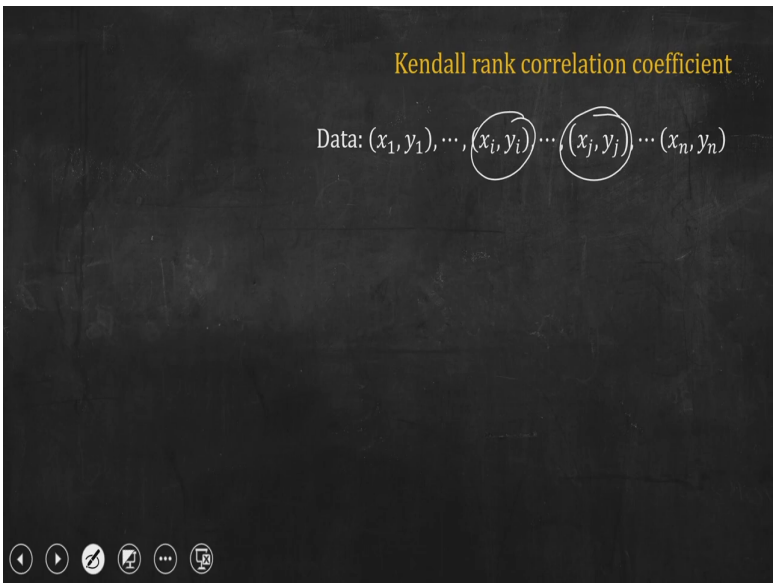
d ← read.csv("data.csv")

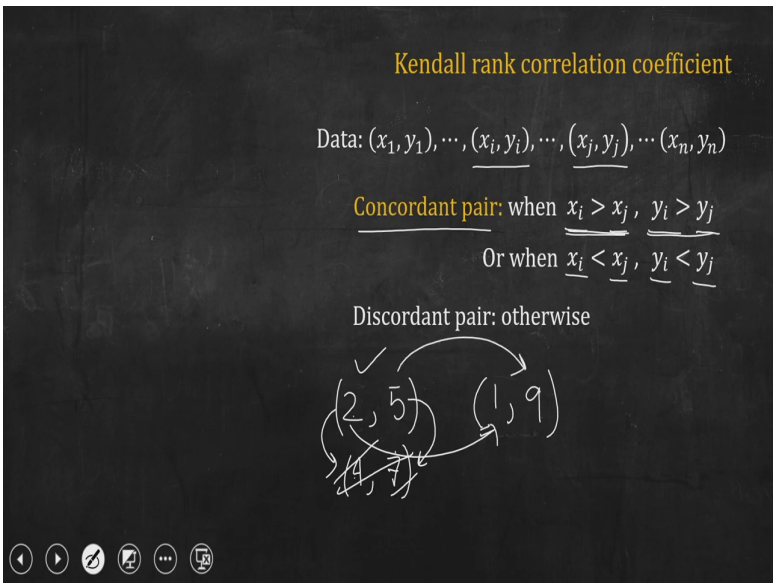
t ← cor(d$v1, d$v2, method = "kendall")

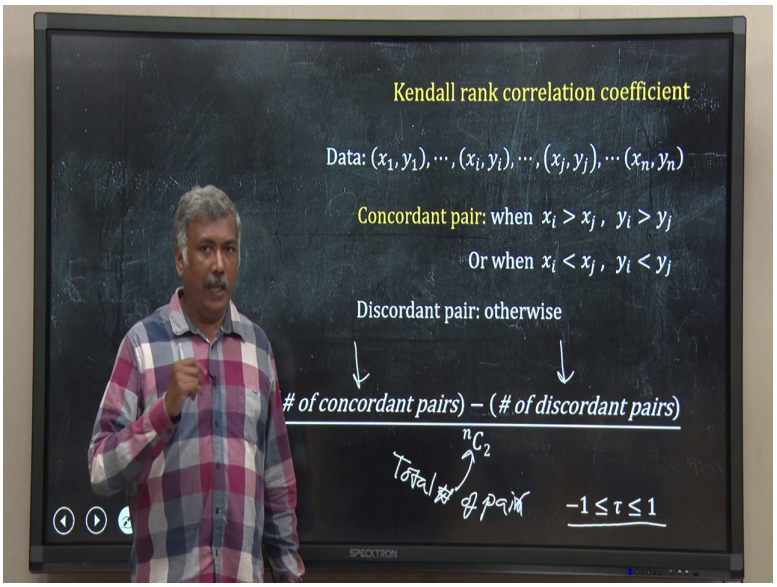
t ← cor(d, method = "kendall")

Second that we discussed now is Spearman and there is another third method which you can use that is called Kendall rank correlation coefficient. When to use it you have to simply define it like method equal to Kendall and then it will calculate the Kendall rank correlation coefficient. Kendall rank correlation coefficient is also a rank correlation coefficient that means it works on rank of the data.

(Refer Slide Time: 31:30)







$$\tau = \frac{(\#\ of\ concordant\ pairs) - (\#\ of\ discordant\ pairs)}{^nC_2}$$

So, let me explain it very briefly without going into details of the mathematics of that. So, suppose this is my data x1, y1, x2, y2, xi, yi these are data. You can write in the two columns, I have written them side by side in one line to make the board bit cleaner. So, now take any two pair suppose xi, yi, xj, yj these two pair you take and now you check out are these two pair concordant or discordant.

Now what do I mean by concordant or discordant? Let me explain, so if a pair is concordant, when I will call a pair concordant? So, suppose I have taken these two pair xi, yi and xj yj, these two pair will be called concordant, when if xi is greater than xj then yi is also greater than yj or when xi if xi is less than xj, yi is also less than yj. Let me give an example. So,

suppose I have one value of x is 2 and the corresponding y is 5, 2, 5 and I have another data point where I have 4 and 7.

So, when 2 has increased to 4, 5 has also increased to 7. So, this relationship is matched, xi is greater than xj, so, 4 is greater than 2. In that case yi is also greater than yj, 7 is also greater than 5. So, this is a concordant pair. When this relationship is broken, let me show you, give an example. Suppose, I have something like this. Suppose, I have 1, 9 forget about this one. Take this one 2, 5 and 1, 9. So, 2 to 1 I have decreased x but in that process y has got increased. So, this is non-concordant pair, this is discordant pair.

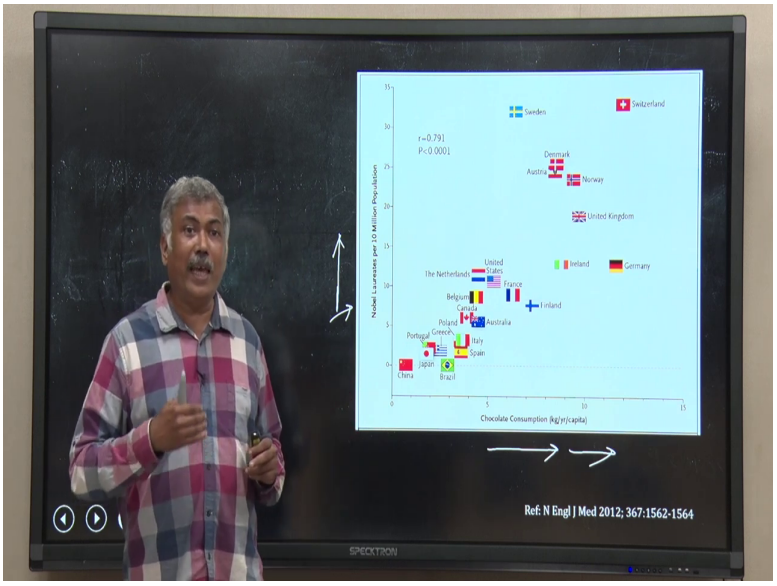
So, what we do? In this whole data set, we count all the concordant pairs, their numbers and the number of discordant pair and then we use the definition or formula of Kendall rank correlation coefficient and that is usually called Kendall's tau and that is equal to number of concordant pair minus number of discordant pair divided by total number of pair, n choose 2 is total number of pair. If I have n number of data sets, then I will have n choose 2 pairs.

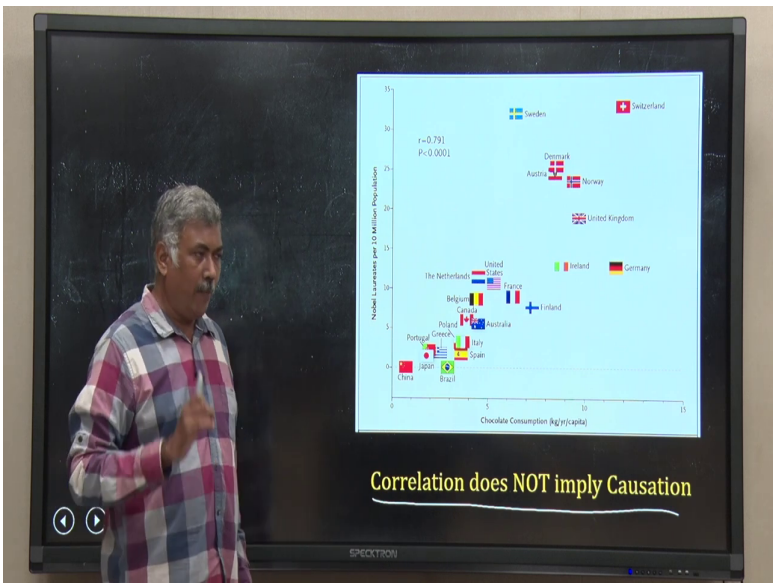
So, I calculate the number of concordant pairs subtract that this number of discordant pair and divide by or normalize by the total number of pair possible in my data set and you can easily show that this tau, Kendall's tau will be varying from plus 1 to minus 1. When there is no discordant pair, it is plus 1, that means every data points are agreeing.

All have the same trend, then obviously my value will be, a tau value will be plus 1, when they have just the opposite thing, everything is discordant. Then it will be minus 1 and when they are equal number then I will be in the middle 0. So, what we have learned and discussed till now we have discussed about three types of correlation coefficient. We started with Pearson correlation coefficient, then we moved into Spearman correlation coefficient and then we moved into the Kendall correlation coefficient.

So, it is very easy to calculate correlation between variables in your data set and even if you do not use r almost all statistical tool and software will have some sort of tool to calculate all these three correlation functions, correlation coefficient and other correlation coefficient which exist and that is where there is a big trap. As it is very easy to calculate, it is very handy and it is very very easy to understand. We very frequently use it and very frequently draw wrong conclusions. Not just in our day to day life but also in scientific articles also.

(Refer Slide Time: 36:02)



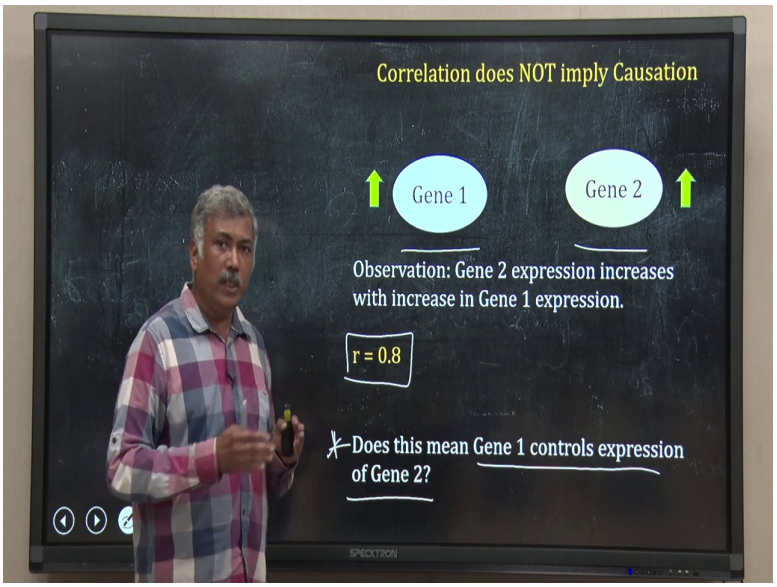


Let me give an example, where we very frequently fall into a trap of correlation. What I have shown here, taken from a scientific journal. In the horizontal axis, we have chocolate consumption per capita of different countries. What it is? Consumption of chocolate per capita of different countries and in the vertical axis what I have? I have number of Nobel Laureates per 10 million population of those countries and the data points are represented by the flag of individual of those countries. Just look at the data, looking at the data it seems there is a linear relation. As chocolate consumption is increasing number of Nobel Laureate in those countries are increasing, is it? And in fact, they have calculated the correlation coefficient also and that is quite a reasonable 0.79 or something by positive, good positive value.

So, now if I have this data and have good positive correlation, can I say that if we all start consuming chocolate then the number of Nobel Laureate in our country will increase? You must be laughing at it. This is absurd. From your common sense, you dont need to dig into the literature for chocolate and all this thing, from your common sense you will tell me that this is a mere coincidence, that there is some sort of correlation positive relation we are seeing between number of Nobel laureate in a country and how much chocolate the population eat in that country. It is a mere coincidence.

So, this type of correlation actually can lead us to a false notion. False conclusion. What we have to remember, as you have just correctly thought of, I will just reframe it in a right sentence. We have to remember that correlations does not imply causation. In this example, there is a mathematical correlation between two data, two variables but that does not mean the chocolate consumption causes Nobel Prizes. So, correlation does not imply causation, as this is a very common mistake in day to day life as well as in scientific literature. Let me give few more example to bring home the message very clearly.

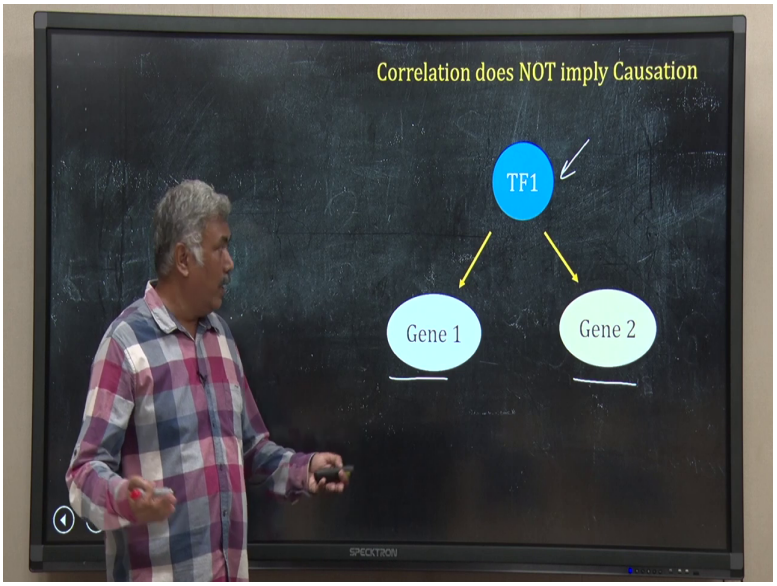
(Refer Slide Time: 38:25)

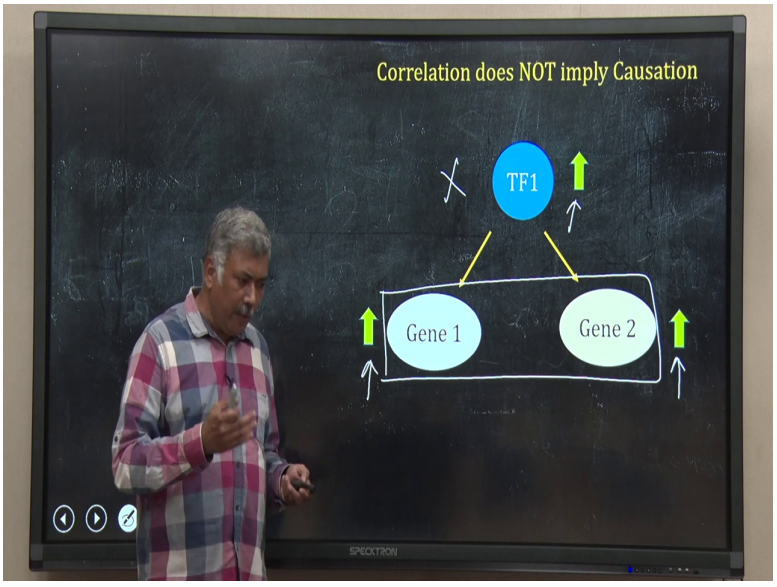


Suppose, I have done an experiment. In that experiment, I have seen every time gene 1 increases, gene 2 also increases and you have a large data set in excel sheet. You have used r or excel to calculate the correlation coefficient and it turns out to be 0.8 positive correlation. So, now I can ask one question, does this means that the gene 1 controls expression of gene 2? That means, I am saying gene 1 is the causal agent of gene 2 expression. Can I conclude this from this correlation coefficient, from this high positive correlation coefficient? It may

be, this can be one hypothesis but I cannot prove that gene 1 controls expression of gene 2 just from this correlation value because there can be another situation.

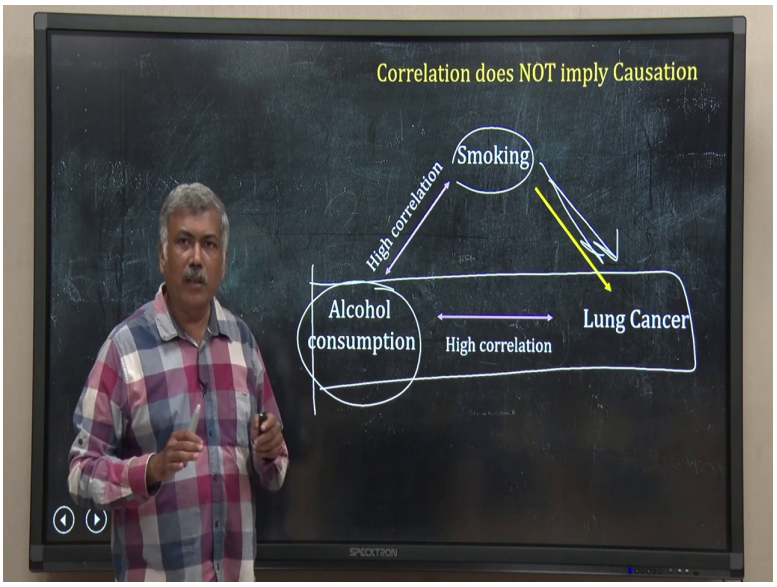
(Refer Slide Time: 39:21)





Imagine there is a transcription factor TF 1 which controls both gene 1 and gene 2 and you do not know that is hidden to you and then when transcription factor one increases both the expression of gene 1 and gene 2 increases. You do not see this one, you see only this part in your experiment and your correlation coefficient calculation says there is a high positive correlation. But maybe the real truth is that both of them are controlled by transcription factor 1 and that is why they have behaved similarly.

(Refer Slide Time: 40:11)

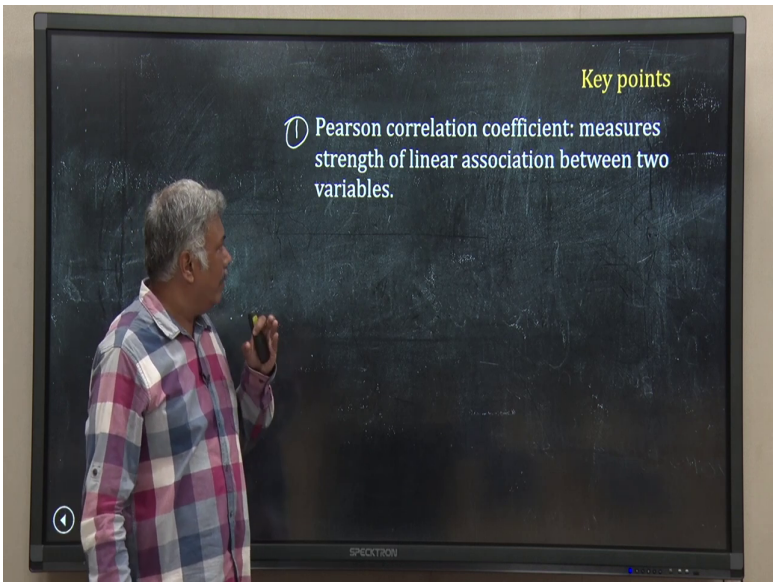


Let me give another example where we may draw a wrong conclusion just from a correlation coefficient value whichever method we use. It does not depend on the method we are using. In fact this has been mislead people earlier. People have seen, if you go into old literature you will see, people have seen that there is a high correlation between alcohol consumption and lung cancer. Lots of epidemiological studies are there old where people have seen there is a positive correlation between alcohol consumption and lung cancer.

Does that mean alcohol consumption causes lung cancer? Actually, as we know now that the main cause for all of these cases possibly is smoking. We know for sure now smoking can cause cancer so there is a causal link between smoking and lung cancer. But at the same time there is a behavioural connection that people who smoke many of them also consume alcohol and many of the people who consume alcohol at the time of consumption of alcohol they smoke.

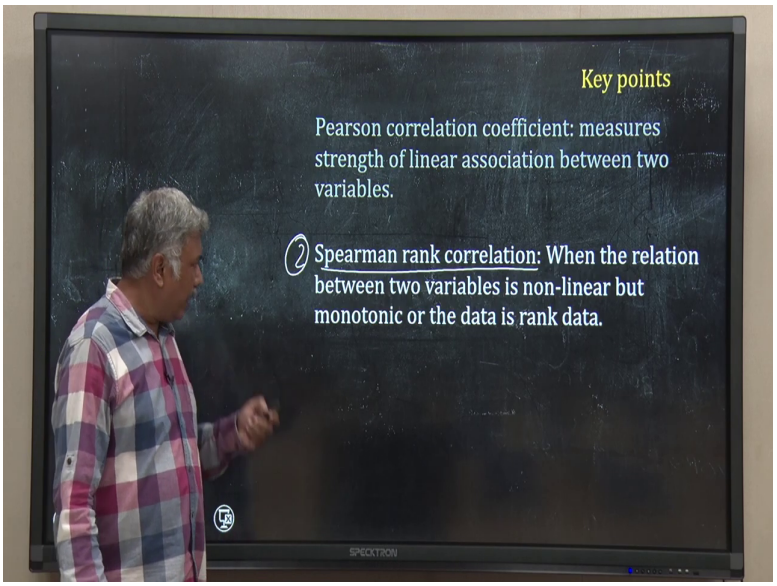
So, that means there is a positive correlation between smoking and alcohol consumption. At the same time smoking causes cancer that is why when you have seen only this part of the data, only alcohol consumption and lung cancer, you have seen a positive correlation. So, alcohol consumption in this case is not the causal agent. It is an associated variable. It also has the same relation with the causal agent thats smoking.

(Refer Slide Time: 41:44)



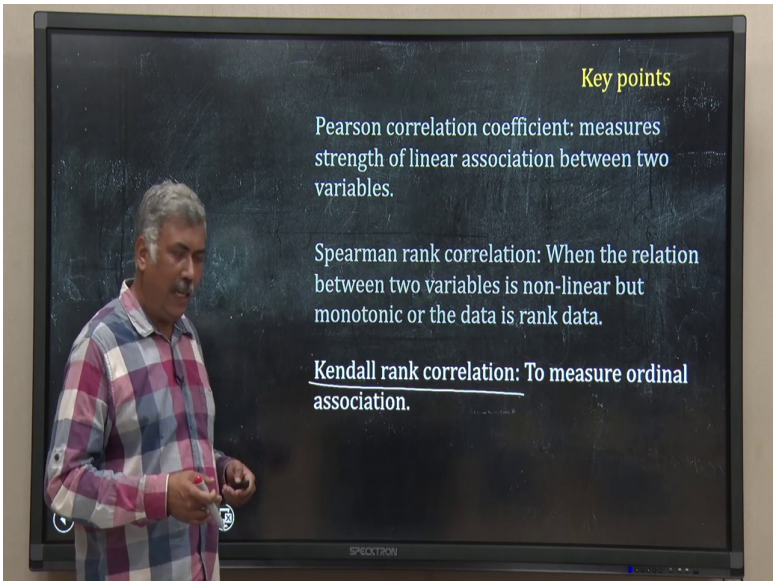
So, that is all for this discussion. Let me jot down what we have learnt in this lecture. We have learnt three type of correlation coefficient. First, we have learned about Pearson correlation coefficient. Pearson correlation coefficient measure the strength of linear association between two variables and it varies from minus 1 to plus 1 and I have explained that. And this is not good if there is a non-linear relation between the variables in your data set. So, you have to be very careful to use this.

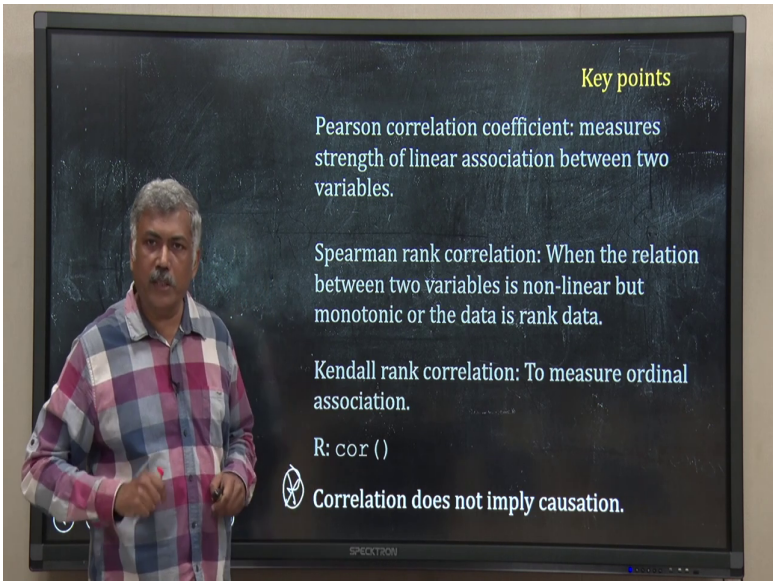
(Refer Slide Time: 42:22)



When there is a monotonic non-linear relationship or my data is category or rank data, then it is better to use the second one, the Spearman's rank correlation. It also will vary from plus 1 to minus 1.

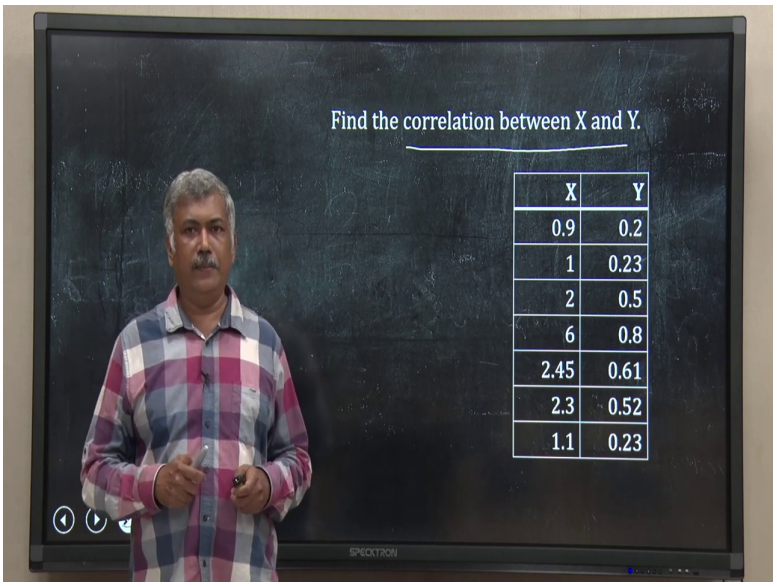
(Refer Slide Time: 42:31)





The third one that we have learned is another ranked correlation coefficient that is called Kendall rank correlation coefficient and all these three thing you can actually very easily calculate using the cor function in r and the last critical point that I discuss in this lecture and we should remind every day to us that correlation in my data does not imply causation.

(Refer Slide Time: 43:02)



X	Y
0.9	0.2
1	0.23
2	0.5
6	0.8
2.45	0.61
2.3	0.52
1.1	0.23

That is all for this lecture. Let me end this lecture with a home task. I have given a data set, two variables X and Y you have to find the correlation between X and Y. Now, notice one thing in this case I have not said what type of correlation you will measure. So, when I do not know what type of correlation I have to measure, there is a way to proceed.

That is I should try to visualize, if possible I should try to visualize the data and then I will choose the right correlation coefficient that I should use for this data set. So, you should also proceed that way try to visualize the data and then decide which one you will use, Pearson, Spearman or Kendall rank correlation. Try to do that using the cor function in r. We will meet again in the next lecture, till then happy learning.