Essentials of Data Science with R Software- 1
Professor Shalabh
Department of Mathematics and Statistics
Indian Institute of Technology, Kanpur
Lecture 47
Covariance and Correlation

Hello friends, welcome to the course Essentials of Data Science with R Software- 1, in which we are trying to understand the basic concept of probability theory and statistical inference. So, up to now you have learnt many things univariate random variable, bivariate random variable and I am sure that if I ask you to extend them to the multiple or multivariate cases it is not problem, unless and until you are clear with the bivariate.

Now, the question is this why I am trying to do all those things, why I have computed or why I have explained you the concept of bivariate random variable, what I am really going to do with this. So, now today in this lecture I am going to talk about it, you see when I discuss the concept of univariate random variable, I explain you that there is only one variable which is going to affect the outcome, now when I say there is a bivariate case that means there are two random variables, so we assume that they are jointly going to affect the outcome, so when they are jointly affect the outcome, so there can be effect of random variable 1, random variable 2 as well as their joint effect.

Now, when you are trying to talk about the joint effect, what does this mean, just by saying, just by looking you cannot understand it, unless and until you quantify it and now when you are trying to talk about the joint effect, that means x and y are jointly going to affect the outcome, that means there is some relationship between x and y. Now, this relationship can be linear in nature or that can be non-linear in nature, so now there are many, many questions which crop up before us, that how to study this relationship, how to quantify the degree of relationship.

So, now we are more interested that we have a bivariate setup where we have two random variables x and y and both are going to affect each other, like I say height and weight of children, when the height increases then weight also increases, when there is a increase in the weight then under general circumstances, general condition the height also increases. Similarly, if you try to see the relationship between the yield of a crop and the quantity of fertilizer, as you try to increase the quantity of the fertilizer up to certain extent the yield of the crop also increases.

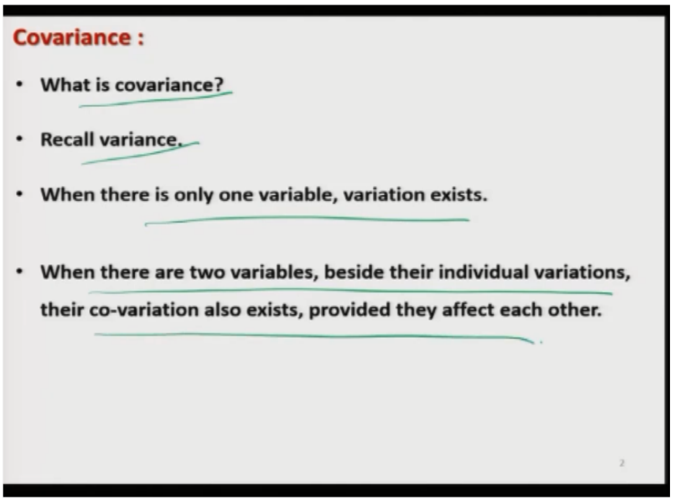
So, now I want to learn about this relationship, so the first thing is this how is this relationship and how to measure it in a quantitative way and for that we are going to use the concept that we have understood in the case of bivariate distribution and remember we also have learnt that how to take the expectation when we have a bivariate probability function and now we are going to use them. So, you know that if you have got a bivariate distribution you can understand, you can learn the joint distribution as well as their marginal distribution also.

So, now using those things we are going to extend the definition of variance to define a quantitative measure for the measurement of the degree of linear relationship between x and y, yes, there can be two types of relationship linear and non-linear but here in this lecture we are going to concentrate on the degree of linear relationship.

So, we are going to consider that there are two random variable x and y which are linearly related, so if you try to understand or in case if you try to recall the concept of variance, what was that? It is trying to measure the variability. Now, if there are two variables, can you define that what is their covariability, variable x has its variation, variable y has its variation but now they are dependent on each other, so they have there some co variation also co variation, that means when the variation is happening together, that is the concept of covariance.

So, we are going to introduce this concept of covariance and then we will try to implement it to define the correlation coefficient which is going to measure the degree of linear relationship between two variables and after that we will see how to use it, how to implement it.

(Refer Slide Time: 5:43)

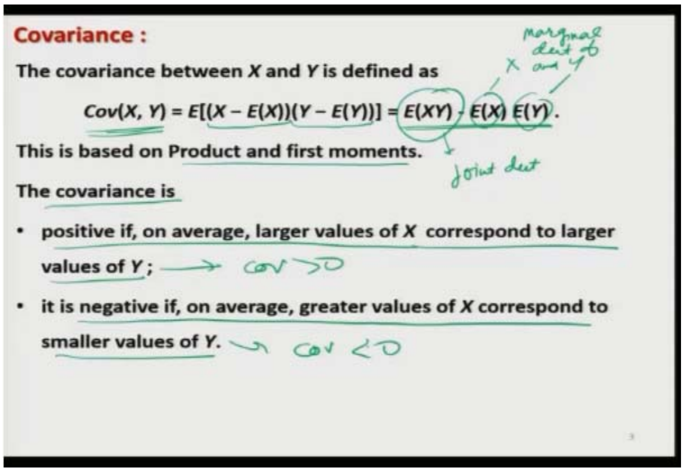


Covariance :

- What is covariance?
- Recall variance.
- When there is only one variable, variation exists.
- When there are two variables, beside their individual variations, their co-variation also exists, provided they affect each other.

So, let us begin our lecture. So, now the first question comes here what is covariance? So you can recall the variance, when there is only one variable only the variation exists, when there are two variable beside their individual variation their co-variation also exists provided they affect each other they are not independent.

(Refer Slide Time: 6:06)



Covariance :

The covariance between X and Y is defined as

$$Cov(X, Y) = E[(X - E(X))(Y - E(Y))] = E(XY) - E(X)\,E(Y).$$

marginal dist of X and Y

joint dist

This is based on Product and first moments.

The covariance is

- positive if, on average, larger values of X correspond to larger values of Y; → $cov > 0$
- it is negative if, on average, greater values of X correspond to smaller values of Y. → $cov < 0$

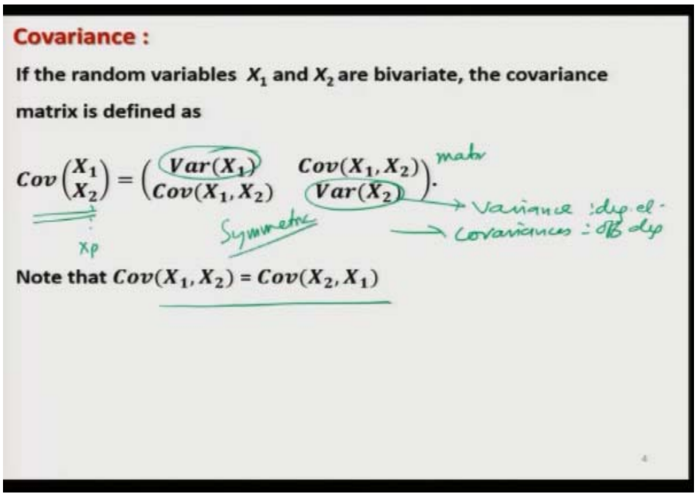
So, this covariance between X and Y is defined as $Cov(X, Y) = E[(X - E(X))(Y - E(Y))]$ they are not a variable like you had done earlier that Y - m type of thing that you are trying to measure it around the mean, yes, and if you try to open this bracket and if you simply try to simplify it you will get here $E(XY)$ - $E(X)$ $E(Y)$.

So, that means if you want to compute it you are going to obtain the E(XY) using the joint distribution and in order to compute this E(X) and E(Y) you are going to use the marginal distribution of your X and your Y. So, this is how you can compute this covariance between X and Y.

So, this is actually based on the product and first moments and this covariance is actually going to indicate the degree of the co-variation also, here this covariance is going to indicate the direction of the co-variation also, what do you mean by this? That the direction can be positive or negative, that means if a variable is increasing then the other variable is also increasing or vice versa, that if one variable is increasing then other variable is decreasing.

So, based on that we can have two direction, positive direction and negative direction, so the covariance is positive if on an average the larger values of X correspond to the larger values of Y and this will be indicated by that the value of covariance is positive and this is negative, this means if on an average greater value of X correspond to smaller values of Y and in this case this covariance is going to be negative.

(Refer Slide Time: 8:20)



Covariance :

If the random variables X_1 and X_2 are bivariate, the covariance matrix is defined as

$$\underline{\underline{Cov\begin{pmatrix} X_1 \\ X_2 \end{pmatrix}}} = \begin{pmatrix} \boxed{Var(X_1)} & Cov(X_1, X_2) \\ Cov(X_1, X_2) & \boxed{Var(X_2)} \end{pmatrix}.^{\text{matr}}$$

$\rightarrow$ Variance : diag. el.

$\rightarrow$ Covariances : off diag

xp

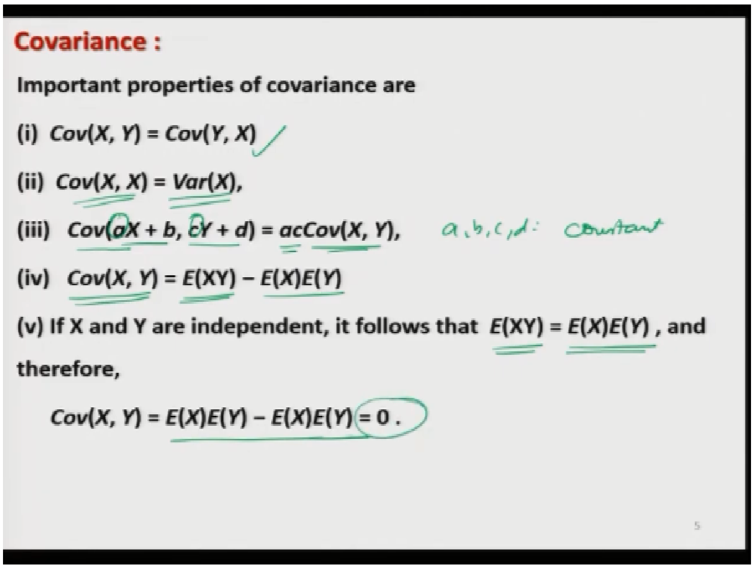
Symmetric

Note that $\underline{Cov(X_1, X_2) = Cov(X_2, X_1)}$

So, now in case if you have suppose two random variables X_1 and X_2 they are bivariate, so one option here is that you can define the covariance matrix, this covariance matrix is defined here like this covariance between X_1 and X_2. So, this is a matrix here where the diagonal elements are going to indicate the variances on the diagonal elements and there will be co-variances on the off-diagonal elements.

Well, I have taken here the symbol X_1 and X_2 instead of X and Y because this can be extended for $X_1, X_2, \ldots, X_p$ exactly in the same way that on the diagonal elements you are going to get the variances of all the variables and in the off-diagonal elements you will get the covariance of the respective random variables. And note that that covariance between X_1 and X_2 is the same as covariance between X_2 and X_1, so this matrix is going to be symmetric, that is obvious, means if height is affecting the weight, then weight is also affecting the height.

(Refer Slide Time: 9:33)



Covariance :

Important properties of covariance are

(i) $Cov(X, Y) = Cov(Y, X)$ ✓

(ii) $Cov(X, X) = Var(X)$,

(iii) $Cov(aX + b, cY + d) = acCov(X, Y)$, a, b, c, d: Constant

(iv) $Cov(X, Y) = E(XY) - E(X)E(Y)$

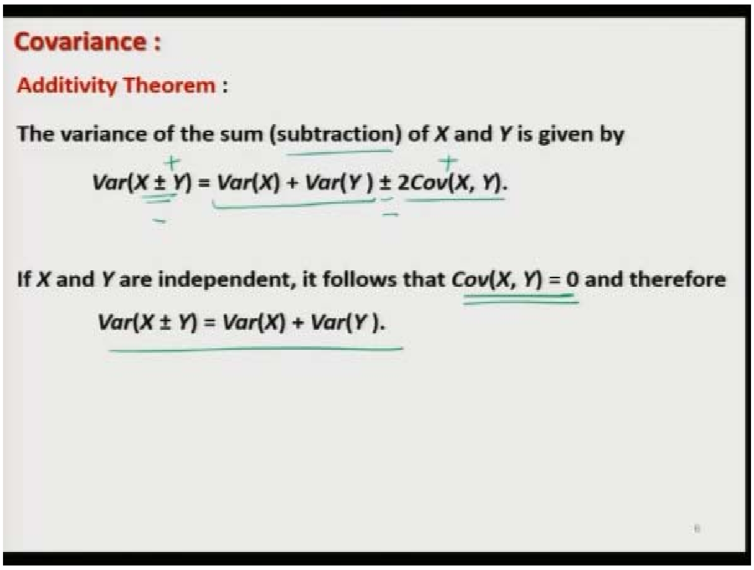
(v) If X and Y are independent, it follows that $E(XY) = E(X)E(Y)$, and therefore,

$$Cov(X, Y) = E(X)E(Y) - E(X)E(Y) = 0 .$$

Now, there are some important property that you have to keep in mind, that covariance between X and Y is the same as covariance between Y and X and if you try to take here covariance between X and X that is nothing but your variance of X, and in case if you try to find out the value of covariance between (aX + b) and (cY + d) where this a, b, c, d they are some constant values, then this is the same as a into c into covariance between X and Y, so you have to consider only the coefficient of X and Y that is what you have to keep in mind and the result that I already shown you that covariance between X and Y can be expressed as E(XY) - E(X) E(Y).

And if X and Y are independent, then obviously, you know from the result that we had done for the stochastic independent that E(XY) can be expressed as E(X) E(Y) and therefore, this covariance will become simply here expected value of X into expected value of Y minus expected value of X into expected value Y which is 0, so these are very important property that you can keep in mind but then I would like to give you here some more results which will be useful for finding out different types of expression.

(Refer Slide Time: 11:00)



Covariance :

Additivity Theorem :

The variance of the sum (subtraction) of X and Y is given by

$$Var(X \overset{+}{\underset{-}{\pm}} Y) = Var(X) + Var(Y) \overset{+}{\underset{-}{\pm}} 2Cov(X, Y).$$

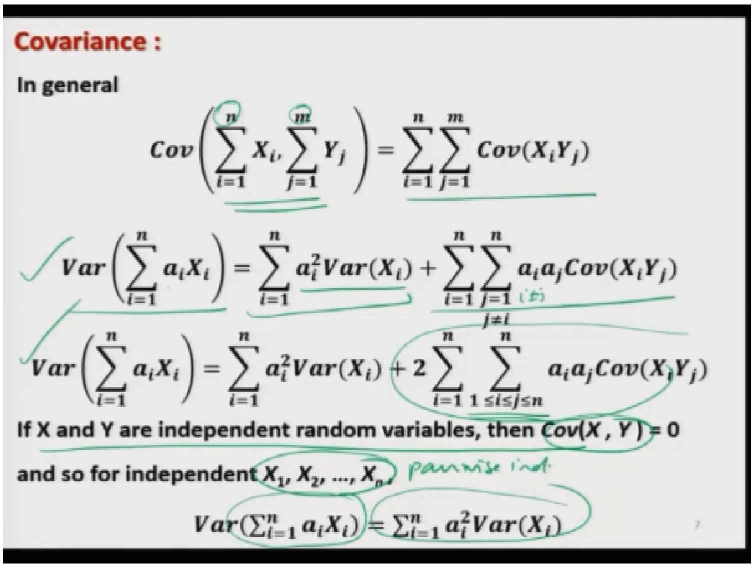
If X and Y are independent, it follows that $Cov(X, Y) = 0$ and therefore

$$Var(X \pm Y) = Var(X) + Var(Y).$$

So, the variance of the Var(X ± Y) = Var(X) + Var(Y) ± 2Cov(X, Y). So, if you are taking here plus it will be here plus, if you are taking here minus then it is going to be here minus and in case if the X and Y are independent then obviously this covariance term is going to be 0 and therefore, the Var(X ± Y) will simply be the Var(X) + Var(Y).

And if you remember once when I was trying to find out the variance of the sample mean at that time I had used this property, that since we have got the random sample, so all the observations were independent, so when we are trying to find out their covariance that was 0.

(Refer Slide Time: 11:54)



Covariance :

In general

$$Cov\left(\sum_{i=1}^{n} X_i, \sum_{j=1}^{m} Y_j\right) = \sum_{i=1}^{n}\sum_{j=1}^{m} Cov(X_iY_j)$$

$$Var\left(\sum_{i=1}^{n} a_iX_i\right) = \sum_{i=1}^{n} a_i^2 Var(X_i) + \sum_{i=1}^{n}\sum_{\substack{j=1 \\ j\neq i}}^{n} a_ia_jCov(X_iY_j)$$

(≠i)

$$Var\left(\sum_{i=1}^{n} a_iX_i\right) = \sum_{i=1}^{n} a_i^2 Var(X_i) + 2\sum_{i=1}^{n}\sum_{1\leq i\leq j\leq n}^{n} a_ia_jCov(X_iY_j)$$

If X and Y are independent random variables, then $Cov(X, Y) = 0$

and so for independent $X_1, X_2, \dots, X_n$, pairwise ind.

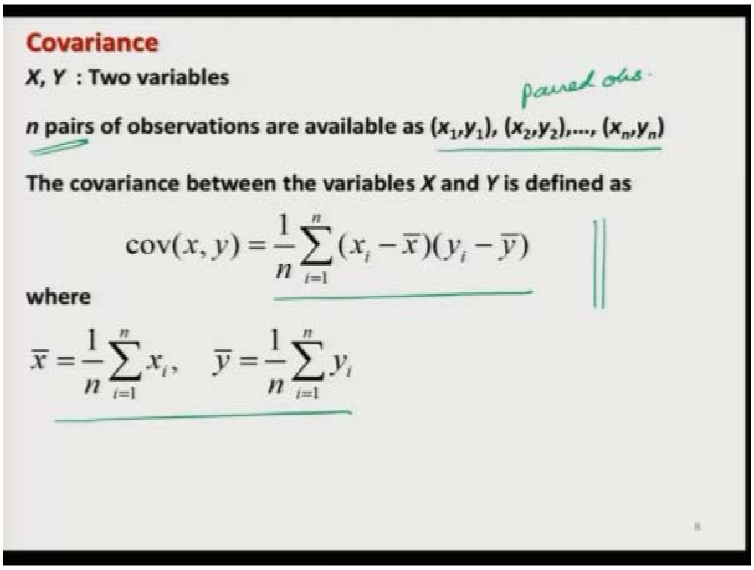
$$Var(\textstyle\sum_{i=1}^{n} a_iX_i) = \sum_{i=1}^{n} a_i^2 Var(X_i)$$

So, now in case if you have to find out the covariance between the linear functions of X and Y like as X_1 plus X_2 plus X_n and Y_1 plus Y_2 plus Y_m, you can see here it is here n and m, then the $Cov(\sum_{i=1}^{n} X_i, \sum_{j=1}^{m} Y_j) = \sum_{i=1}^{n}\sum_{j=1}^{m} Cov(X_iY_j)$ and similarly if you want to find out here the variance of this linear function of this Xi that is something like a1X1 plus a2X2 plus anXn then this can be written here as $\sum_{i=1}^{n} a_i^2 Var(X_i)$ + $\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n} a_ia_jCov(X_iY_j)$.

So, and the same expression can also be written if you want to write down here that 1 less than equal to i less than equal to j greater than equal to n then this can be written as here like this thing twice of this summation, so this expression and this question they are the same thing, so and definitely if all this Xi's are independent then this covariance term will become 0 and we can write down here that variance of the summation aiXi will be simply sum of the variances that is summation i goes from 1 to n ai square variance of Xi.

So, and this can be obtained because we are assuming that X and Y are independent so covariance between X and Y will be 0 but when I am trying to say that all the variables X_1, X_2,.., X_n they are also independent that means they are pair wise independent. If they are pair wise independent then they will they will all be mutually independent, that is the property that we already had talked earlier.

(Refer Slide Time: 13:56)



Covariance

X, Y **: Two variables**

paired obs.

n **pairs of observations are available as** $(x_1,y_1), (x_2,y_2),\ldots, (x_n,y_n)$

The covariance between the variables *X* and *Y* is defined as

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

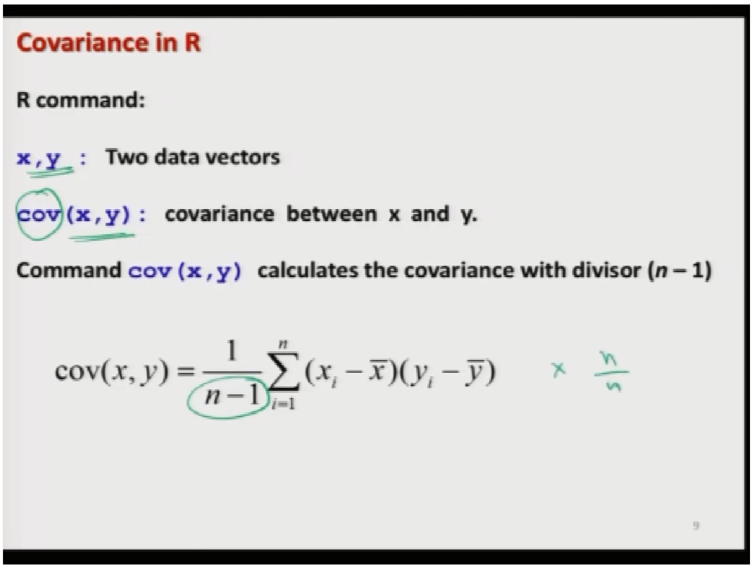
where

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

8

So, now the question here is suppose you have got some data sets suppose, you have got here n pairs of observations which are like (X_1, Y_1), (X_2, Y_2), (X_n, Y_n), so these are your paired observations. Now, in this case if you want to compute the sample base value of a covariance between X and Y this can be defined here as say $\mathrm{cov}(x,y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$ where $\bar{x}$ and $\bar{y}$ are the arithmetic means of the observations on x_i's and y_i's. So, this is how we try to compute this covariance on the basis of a given set of data and inside the R software also it does the same thing that is why it is important for you to understand it.

(Refer Slide Time: 14:50)



Covariance in R

R command:

x , y : Two data vectors

cov(x , y) : covariance between x and y.

Command cov (x , y) calculates the covariance with divisor ($n - 1$)

$$\text{cov}(x, y) = \frac{1}{\boxed{n-1}}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \quad \times \frac{n}{n}$$

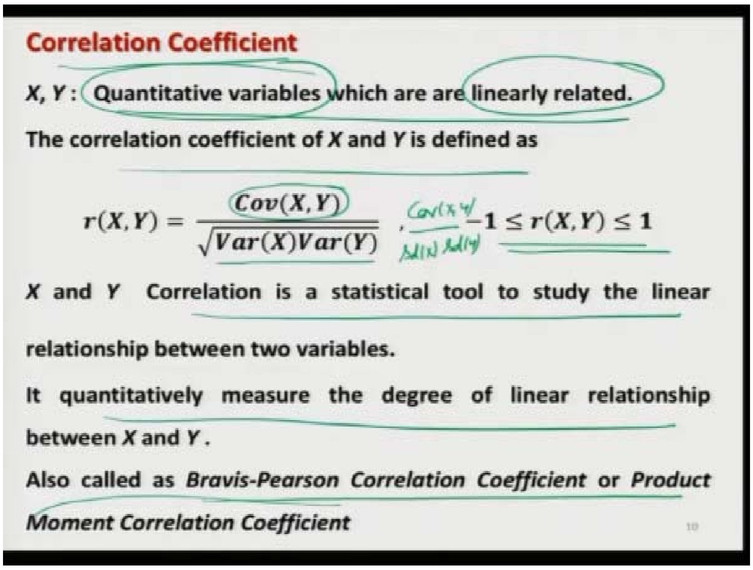
9

Now, in case if you want to compute this covariance inside the R software then the simple command here is cov and then if you have here the data is represented in the form of two data vectors here x and y then the covariance between x and y is written as cov inside the parenthesis (x, y).

So, that will compute the value of covariance between x and y but you have to remember one thing, when you are trying to compute this covariance inside the R software then here the divisor here is n - 1, not n and this has the same interpretation what I had given you when I was explaining you how to compute the variance. In variance also we had defined as the of $\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$ but in the R software it was trying to compute 1/(n- 1).

So, I will not repeat here the discussion but whatever you had learnt at that time and whatever I had explained you at that time that will remain valid here also. And then if you want to really find out the covariance which has the factor 1 upon n that you can obtain by multiplying n divided by here n upon n, so that is not a difficult thing.

(Refer Slide Time: 16:10)



Correlation Coefficient

X, *Y* : Quantitative variables which are are linearly related.

The correlation coefficient of *X* and *Y* is defined as

$$r(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}, \frac{Cov(X,Y)}{sd(X)\, sd(Y)} -1 \leq r(X,Y) \leq 1$$

X and *Y* Correlation is a statistical tool to study the linear relationship between two variables.

It quantitatively measure the degree of linear relationship between *X* and *Y* .

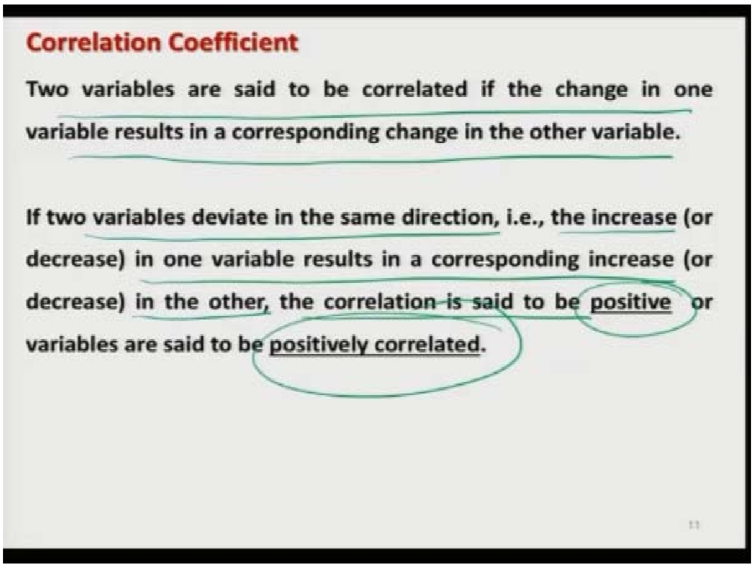
Also called as *Bravis-Pearson Correlation Coefficient* or *Product Moment Correlation Coefficient*

Now, based on this we try to define here a very important quantity correlation coefficient. So, suppose X and Y are quantitative variable which are linearly related, yes, that is very important many people forget this assumption and we are assuming that X and Y are linearly related that is very important, many people forget this thing that the relationship has to be only linear, linear, linear and only linear.

In this case the correlation coefficient between X and Y is defined as the $r(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$ or you can write down here this covariance between XY and divided by standard deviation of X into standard deviation of Y, whatever you want to write and the value of this correlation coefficient lies between - 1 and plus 1, yes, that can also be proved mathematically without any difficulty but anyway I will just inform you here.

So, this correlation coefficient between X and Y is a statistical tool that helps us in studying the linear relationship between the two variable. We can quantitatively measure the degree of linear relationship between X and Y, I will try to show you through the graphical things also and this is also called as Bravis-Pearson Correlation Coefficient or the Product Moment Correlation Coefficient, these are the different names and I am informing you here because when we are trying to compute it in the R software we have to give these options.

(Refer Slide Time: 18:06)



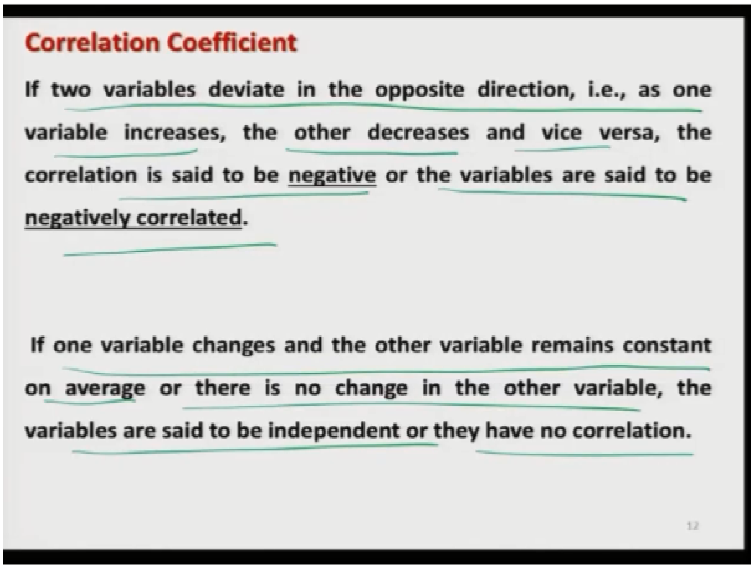
Correlation Coefficient

Two variables are said to be correlated if the change in one variable results in a corresponding change in the other variable.

If two variables deviate in the same direction, i.e., the increase (or decrease) in one variable results in a corresponding increase (or decrease) in the other, the correlation is said to be positive or variables are said to be positively correlated.

Now, what are the interpretations of this correlation coefficients? Two variables are said to be correlated if the change in one variable results in the corresponding change in the other variable. If the two variable they deviate in the same direction, that is the increase or decrease, if one variable increases other also increases or if one variable decreases other variable also decreases. So, if there are two variables, they deviate in the same direction that is the increase or decrease in, increase or decrease in the one variable results in the corresponding increase or decrease in the other variable, then we say that the correlation is positive and the variables are positively correlated.

(Refer Slide Time: 18:53)



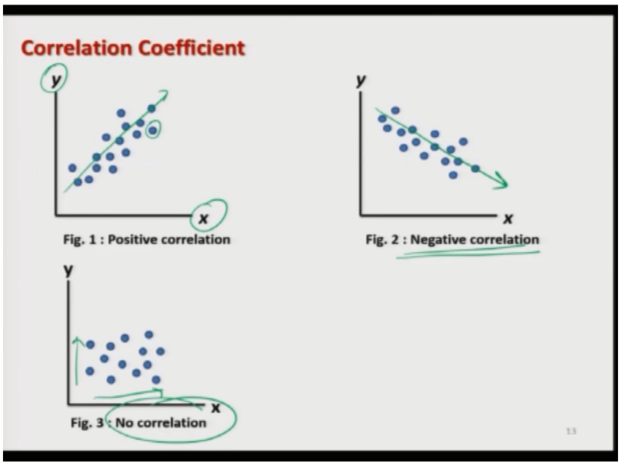
Correlation Coefficient

If two variables deviate in the opposite direction, i.e., as one variable increases, the other decreases and vice versa, the correlation is said to be negative or the variables are said to be negatively correlated.

If one variable changes and the other variable remains constant on average or there is no change in the other variable, the variables are said to be independent or they have no correlation.

And similarly, we can also define the negative correlation and negatively correlated variables. So, if the two variables deviate in the opposite direction, that is as one variable increases the other decreases and vice versa, the correlation is said to be negative or the variables are said to be negatively correlated. If one variable changes and the other variable remain constant that means there is no change in other variable on an average, then what we can say that the variables are independent or they have no correlation. So, now you have three-situations- positive correlation, negative correlation and no correlation.

(Refer Slide Time: 19:39)



Correlation Coefficient

Fig. 1 : Positive correlation

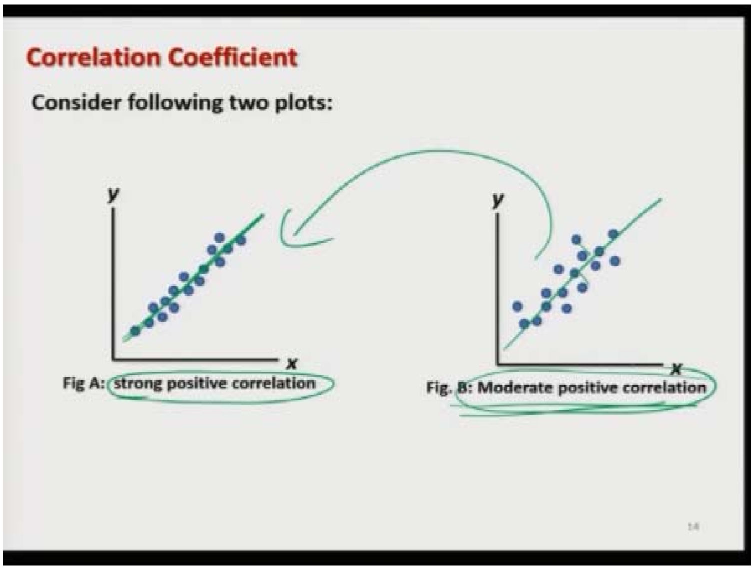
Fig. 2 : Negative correlation

Fig. 3 : No correlation

So, how these things are going to look like let me try to show you by this simple graphics. So, you can see here we have here two variables X and Y on which we have got some observations and we have plotted them. So, these dots are going to indicate the paired observations, so you can see here that X values are increasing Y's are also increasing in this direction, so we can say that there is a positive correlation.

And similarly, you can see here that if the values of X are increasing the Y's are decreasing, so in this type of picture we try to say that there is a negative correlation and if X is increasing then what is the change in Y is not clear, then we say that simply there is no correlation. So, this is how we try to look graphically on the aspect of correlation coefficient on the basis of given set of data.

(Refer Slide Time: 20:39)



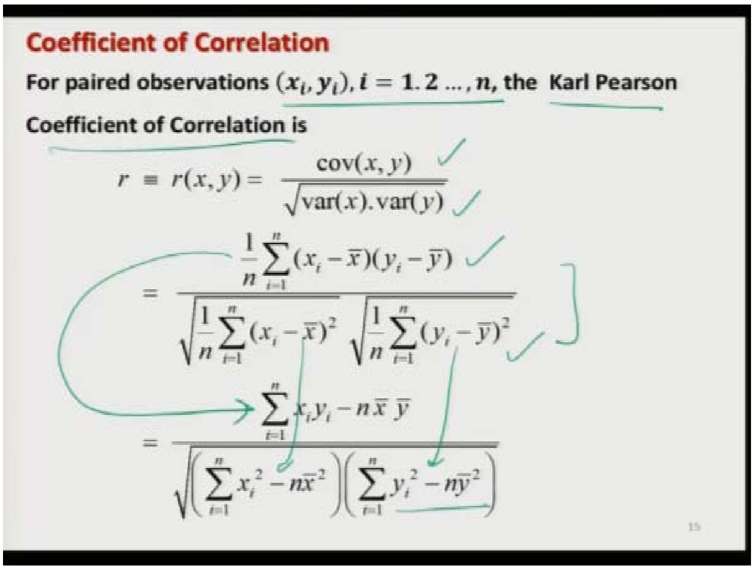
Correlation Coefficient

Consider following two plots:

So, for example, now in case if you try to consider these two pictures, this will give you the idea of the degree, degree of correlation coefficient, you can see here that in the graph number one, here the points are very close to this line and where is in the graph number two here these points are quite a way, means in comparison to this first picture, first graph.

So, in this case if I say that there is a strong positive correlation that most of the points are lying very close to the line then we say that there is a strong positive correlation and in comparison to this in this picture you can say that here the points are lying close to the line but they are not as close as in the first picture, so we say that there is a moderate positive correlation and the same thing can be done for the negative correlation also.

(Refer Slide Time: 21:32)



Coefficient of Correlation

For paired observations $(x_i, y_i), i = 1.2 \ldots, n,$ the Karl Pearson Coefficient of Correlation is

$$r \equiv r(x,y) = \frac{\text{cov}(x,y)}{\sqrt{\text{var}(x).\text{var}(y)}}$$

$$= \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

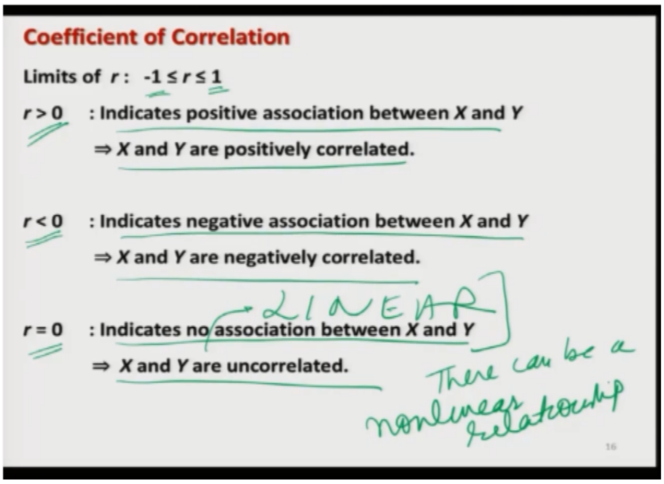
$$= \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\,\bar{y}}{\sqrt{\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)\left(\sum_{i=1}^{n} y_i^2 - n\bar{y}^2\right)}}$$

15

So, now in case if you try to see that how are we going to compute it on the basis of given set of data, then suppose we have here small n pairs of observations on the two variables x and y then the correlation coefficient is defined like this and this is also called as Karl Pearson coefficient of correlation, it is defined here as a covariance between x and y divided by square root of variance of x and variance of y.

Now, if you simply try to substitute the values of covariance and the variance of x and y you get here this expression and you know that this expression can be further solved, this covariance can be written as $\frac{\sum_{i=1}^{n} x_i y_i - n\,\bar{x}\;\bar{y}}{\sqrt{\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)\left(\sum_{i=1}^{n} y_i^2 - n\bar{y}^2\right)}}$. So, this can be computed using this expression.

(Refer Slide Time: 22:28)



Coefficient of Correlation

Limits of r: $-1 \le r \le 1$

$r > 0$: Indicates positive association between X and Y

$\Rightarrow$ X and Y are positively correlated.

$r < 0$: Indicates negative association between X and Y

$\Rightarrow$ X and Y are negatively correlated.

LINEAR

$r = 0$: Indicates no association between X and Y

$\Rightarrow$ X and Y are uncorrelated.

There can be a nonlinear relationship

16

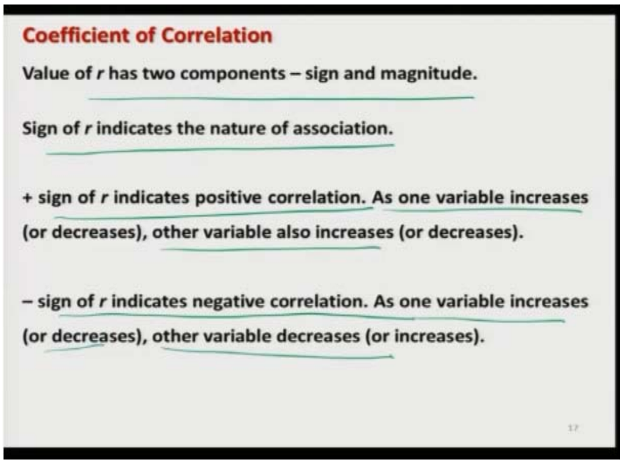
Now, the question comes here how to interpret these values? So, the limits of this correlation coefficients are between - 1 and plus 1, so if r comes out to be greater than 0, suppose you compute this value on the basis of given set of data and if this value comes out to be positive that will indicate that there is a positive association between X and Y that means X and Y are positively correlated.

And similarly, if you compute the value of correlation coefficient and suppose this value comes out to be negative that r is less than 0 then this indicates there is a negative association between X and Y and we say that X and Y are negatively correlated. And similarly, if you try to see, ideally I am saying r equal to 0 but practically if your value of r comes out to be very close to 0 then this will indicate that there is no association between X and Y, and remember one thing this is trying to indicate there is no linear association.

So, X and Y are uncorrelated, once again I will say this is talking of linear relationship, linear association but there can be a non-linear relationship, so this is what you have to always keep in mind, that r equal to 0 is indicating the stochastic independence in the sense that there is no linear relationship between X and Y but there can be a nonlinear relationship between X and Y.

So, many times people do not understand this concept and they try to compute the value of correlation coefficient either they are trying to measure the degree of relationship in a linear or a non-linear relationship, but measuring the degree of non linear relationship using this definition of correlation coefficient is wrong.

(Refer Slide Time: 24:26)



Coefficient of Correlation

Value of *r* has two components – sign and magnitude.

Sign of *r* indicates the nature of association.

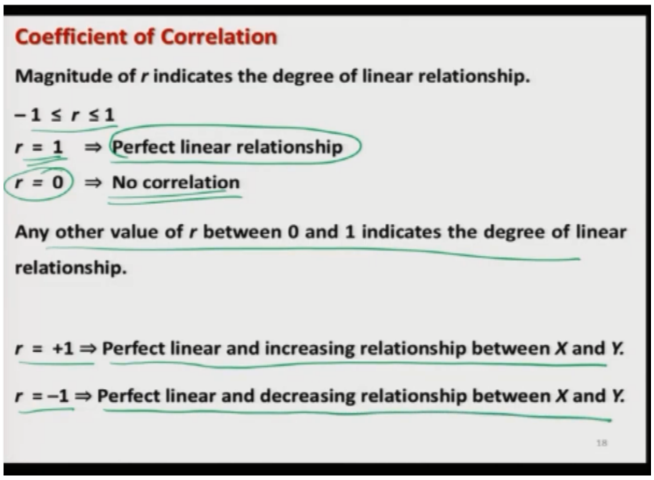
+ sign of *r* indicates positive correlation. As one variable increases (or decreases), other variable also increases (or decreases).

– sign of *r* indicates negative correlation. As one variable increases (or decreases), other variable decreases (or increases).

17

So, now you can see here that the value of r has two components sign and magnitude, the sign of correlation coefficient indicates the nature of association, plus sign of r indicates the positive correlation that is as one variable increases or decreases the other variable also increases or decreases, and the minus sign of r indicates the negative correlation, that means as one variable increases or decreases other variable decreases or increases just opposite.

(Refer Slide Time: 24:58)



Coefficient of Correlation

Magnitude of *r* indicates the degree of linear relationship.

$-1 \leq r \leq 1$

$r = 1$ ⇒ Perfect linear relationship

$r = 0$ ⇒ No correlation

Any other value of *r* between 0 and 1 indicates the degree of linear relationship.

$r = +1$ ⇒ Perfect linear and increasing relationship between *X* and *Y*.

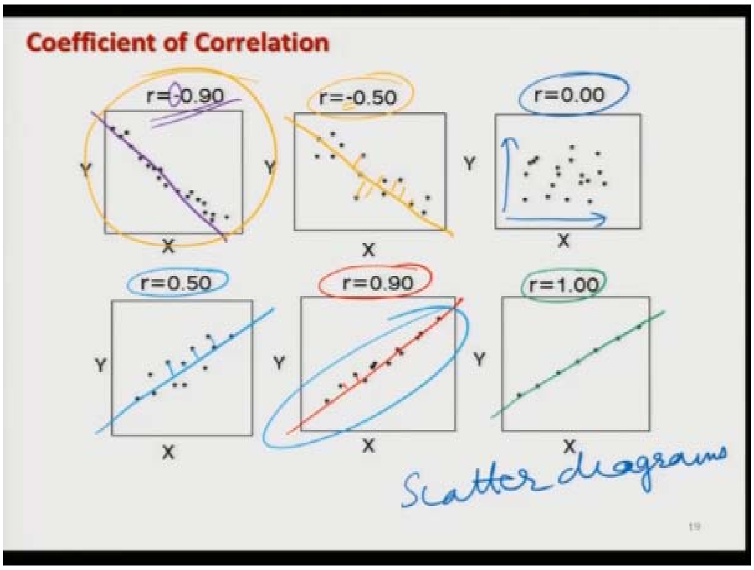
$r = -1$ ⇒ Perfect linear and decreasing relationship between *X* and *Y*.

18

And in case if you try to see what is the interpretation of this exact value. For example, if you try to see that the values of r are lying between - 1 and plus 1, so in case if r is exactly equal to 1 we call this as a perfect linear relationship, and if r is exactly equal to 0 that means there is no correlation. So, obviously any other value of r between 0 and 1 will indicate the degree of linear relationship.

In case if r is equal to plus 1 this will indicate the perfect linear and increasing relationship between X and Y and if r is equal to minus 1 this will indicate the perfect linear but decreasing relationship between X and Y. So this is how we try to interpret this value on the basis of given set of data.

(Refer Slide Time: 25:51)



Coefficient of Correlation

For example, I can show you here some more picture to convince you, you can see here these stars, means stars are actually the some observed values, some or some paired observation on the two random variables X and Y, so just try to observe the pattern of these dots or a stars, you can see here first try to see here in this picture, please try to follow my pen.

In this picture all the points are exactly lying on the straight line, so in this case the r is going to be plus 1. Now, in this picture let me try to change the color of the pen so that you can see it clearly, here is the line but the points are quite away from the line, they are not exactly lying on the line, so in this case you can say that r is close to 0.90, well these are only the indicative values.

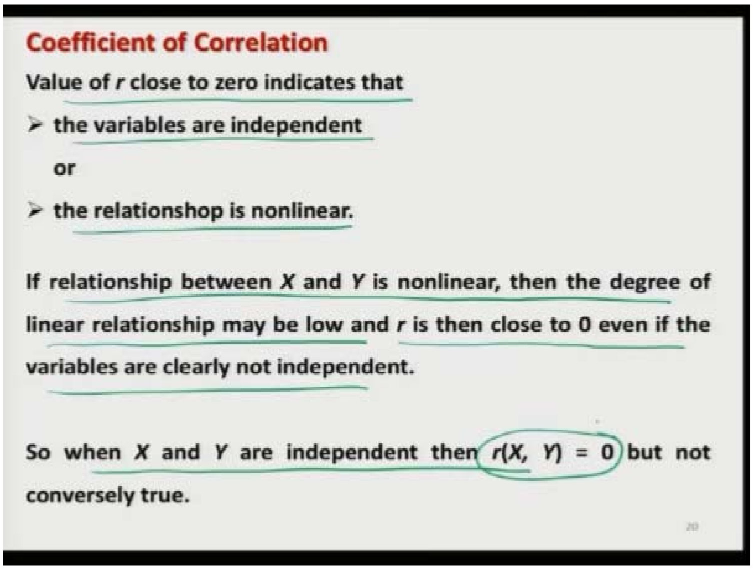
And similarly if you come to this picture here, you can see here or let me try to make it a better picture like this, you can see here that most of the points are close to this line but they are quite a way in comparison to these observations, so in this case the magnitude of r may come out to be 0.5 and the direction will be positive.

Similarly, if you try to now take the first picture here, you can see here in this case as the values of X are increasing the Y's are decreasing, so in this case the sign of r is going to be negative but the points are lying very close to the line, so I can say that the correlation coefficient is close to 0.9.

And similarly, if you try to come to here this picture here you can see here that as X is increasing Y's are decreasing and the points are lying quite away from the line, so in this case the direction is going to be negative but the magnitude is going to be 0.5, 0.5 is trying to indicate that the degree of correlation coefficient is here less in comparison to this data set.

And similarly, if you try to take here this last picture you can see here as X are increasing what is happening to Y? We are not clear there is no trend like this thing, so in this thing you can say that r is here close to 0, so this is how we try to interpret by looking at the scatter diagrams of the observed data that we observe in real data sets, so these are called actually here as a scatter diagrams and you know that you can use you can compute or you can plot them by using the command plot.

(Refer Slide Time: 28:45)



Coefficient of Correlation

Value of *r* close to zero indicates that

- the variables are independent

  or

- the relationshop is nonlinear.

If relationship between *X* and *Y* is nonlinear, then the degree of linear relationship may be low and *r* is then close to 0 even if the variables are clearly not independent.

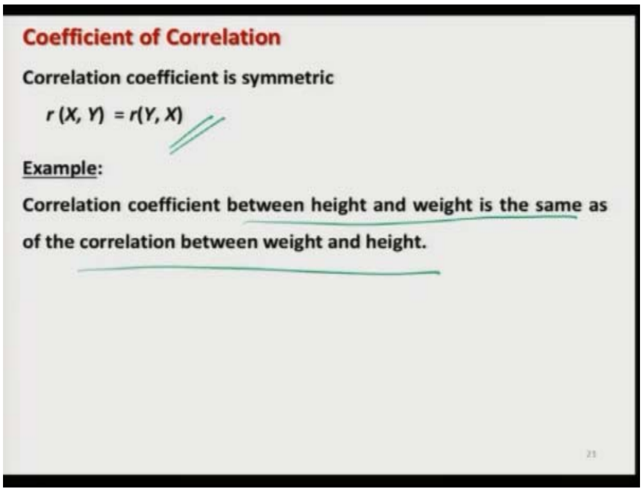
So when *X* and *Y* are independent then $r(X, Y) = 0$ but not conversely true.

20

So, now to conclude this lecture I can see here you have to just keep in mind couple of things, that values of r close to 0 indicate that the variables are independent or the relationship is non-linear, and if the relationship between X and Y is non-linear then the degree of linear relationship may be low and that is why r is close to 0 even if the variables are clearly not independent, they may have some non-linear relationship.

So, the very important rule that you have to keep in mind that when X and Y are independent then correlation coefficient will be equal to 0 but not conversely true. So, this is what you have to keep in mind here.

(Refer Slide Time: 29:36)



Coefficient of Correlation

Correlation coefficient is symmetric

$r(X, Y) = r(Y, X)$

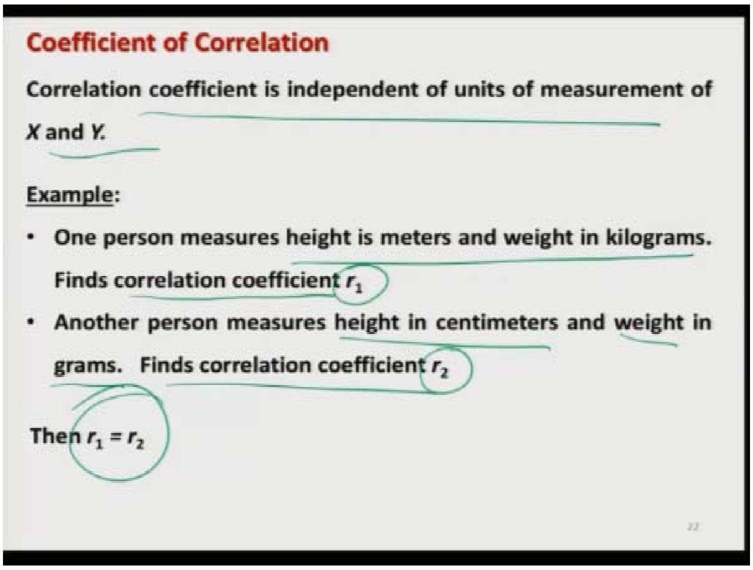
Example:

Correlation coefficient between height and weight is the same as of the correlation between weight and height.

21

And this correlation coefficient is a symmetric, the correlation coefficient between X and Y will be the same as the correlation coefficient between here Y and X and for example, the correlation coefficient between height and weight is the same as the correlation coefficient between weight and height, there is no issue at all.

(Refer Slide Time: 29:55)



Coefficient of Correlation

Correlation coefficient is independent of units of measurement of X and Y.

Example:

- One person measures height is meters and weight in kilograms. Finds correlation coefficient r_1
- Another person measures height in centimeters and weight in grams. Finds correlation coefficient r_2

Then $r_1 = r_2$

22

And one important property that the correlation coefficient is independent of the units of measurement of X and Y. For example, if one person measures the heights in meter and weight in kilogram and suppose the correlation coefficient is found to be r_1 and another person measure the height in centimeter and weight in grams and find out the correlation coefficient suppose the value comes out to be r_2, in both the cases this r_1 and r_2 they are going to be the same.

So, that is an important property and now with this property I come to an end to this lecture and you can see here that was a very interesting lecture a theory as well as application and this is how we are moving forward towards the data science. Now, you can see here by using the concept of correlation coefficient you can very easily find if the two variables on which you have only the data, whether they are linearly related or not just by plotting them and computing the value of r that will give you the magnitude and so on.

But very important point that you always have to keep in mind that r is measuring only the degree of linear relationship, I am repeating again and again please do not forget it number one, number two if the value of correlation coefficient comes out to be close to 0 you can say that X and Y are independent but opposite is not true because if correlation coefficient is close to 0 there is a possibility that there may exist some non-linear relationship also and r is not used to measure the degree of nonlinear relationship.

So, in the next lecture I will try to show you that how are you going to compute these values in the R software, you please try to understand this concept, try to learn this concept and try to settle it in your mind, that will help you, so try to practice it and I will see you in the next lecture till then good bye.