Descriptive Statistics with R Software
Prof. Shalabh
Department of Mathematics and Statistics
Indian Institute of Technology, Kanpur

Lecture – 28
Association of Variables – Correlation Coefficient

Welcome to the lecture on the course Descriptive Statistics with R Software. You may recall that in the earlier lectures we started a discussion on the aspect of association between two variables and more than two variables and we have two types of tools graphical tools and analytical tools. In the last couple of lecture we have discussed the graphical tools, to study the association between or among the variables. Now the next question is how to quantify that association? For example, we have seen in graphics that there can be an association that is strong or that is weak. But now how to quantify where that what value represent a strong or what value represent a weak association?

So, in order to do so we have a concept of correlation coefficient. And this correlation coefficient is a measure of a degree of linear relationship between the two variables, but now when I say the relationship between two variables now this variables can be of different types; they can be continuous, they can be discrete, they can be categorical variables and so on. So, in order to study the quantification of association, we have different tools depending on the nature of variables. We have the concept of correlation coefficient we have the concept of rank correlation we have the concept of contingency tables and so on.

So, we will try to do it in the next couple of lectures. So, in this lecture we are going to discuss about the correlation coefficient and I will be giving you here only the concept and the theory of correlation coefficient means what is the formula what is the structure how you use it what are the interpretations in different situations? And in that next lecture I will show you that how to compute it on a software and how to interpret the graphics and numerical values together. So, let us start our discussion.

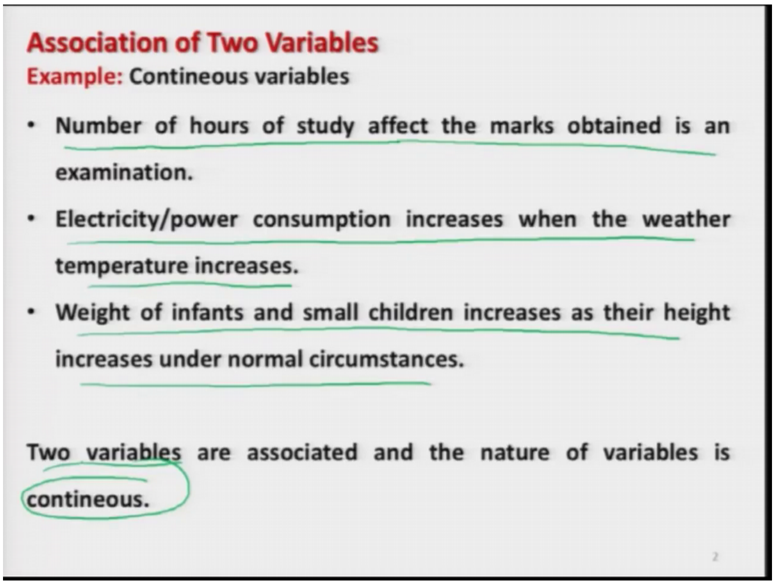
So, we already have understood what is called the association between or among the variables. Now I will try to take several example to explain you that how would you differentiate the association between continuous variable between discrete variables and so on. If I take an example, that we all know that as the number of hours of study increases the students usually obtain more marks. And you can see this number of hours can be expressed as hours minutes and so on, that is essentially the time and marks can also be measured on a continuous scale they can be 70 or 70.50 also.

So, in this case you can see that there is a relationship between, the time is spent on studies and marks obtained in the examination. So, in this case we would like to see how is this association is increasing decreasing and what is there and what is the strength of a relationship? Similarly, if I take another example, we know that during summer times when the weather temperature is high then the consumption of electricity increases people are using air coolers air conditioners and so on.

So, one can feel that as the weather temperature is increasing the consumption of the electricity also increases. But now we would like to verify it on the basis of a given sample of data, we would try to quantify the degree of association and we would like to see that how to say whether the association is a strong moderate or weak and how to see whether the trend is increasing or decreasing and how to quantify it alright. Similarly, if I take another example we know that for the small babies and children, as their age is increasing their weight also increases up to a certain age means after certain age the height and weight both stabilizes.

So, now I can say that as the age of those babies are increasing, their height and also their weight they increase. So, there is a association between the two variables. So, now in these cases the variables are continuous.

(Refer Slide Time: 05:26)



Association of Two Variables

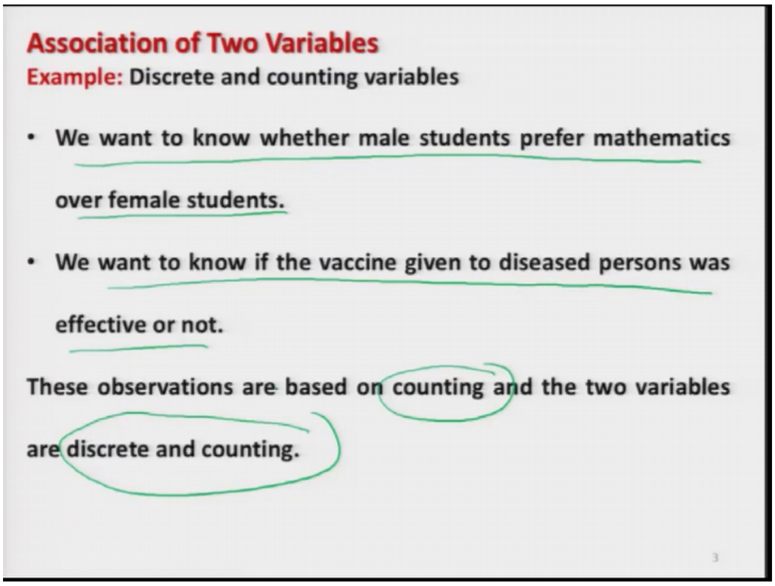
Example: Contineous variables

- Number of hours of study affect the marks obtained is an examination.
- Electricity/power consumption increases when the weather temperature increases.
- Weight of infants and small children increases as their height increases under normal circumstances.

Two variables are associated and the nature of variables is contineous.

So, now we have noticed that in case if the numbers of hours of study, the increase the marks obtained by the students are also increasing. So, number of hours of study they affect the marks obtained in an examination similarly, the power consumption or say electricity consumption increases when the weather temperature increases weight of infants and small children increases as their height increases and their normal circumstances. So, in this case you can see and you can observe that the two variables are continuous in nature. So, the question is how to quantify this association.

(Refer Slide Time: 06:04)



Association of Two Variables

Example: Discrete and counting variables

- We want to know whether male students prefer mathematics over female students.
- We want to know if the vaccine given to diseased persons was effective or not.

These observations are based on counting and the two variables are discrete and counting.

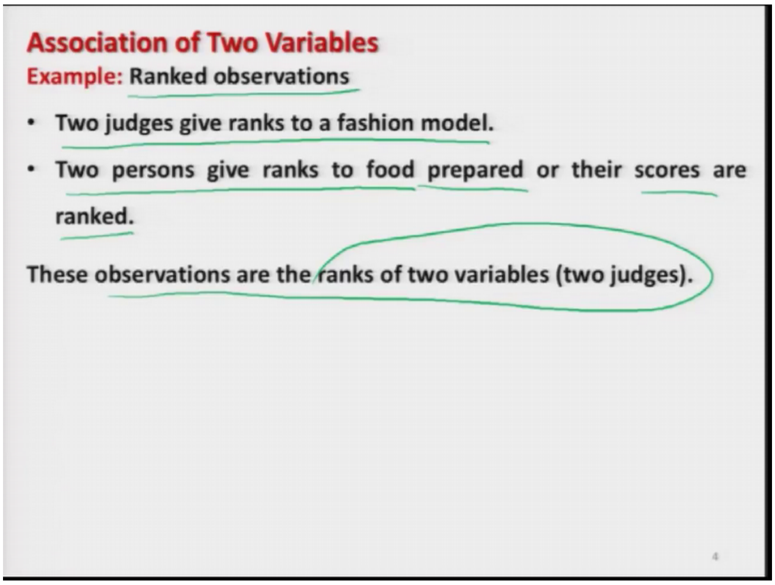
Similarly, if I try to take here another example, in this example I will try to consider the variable which are discrete; their values are obtained only at point right. For example, if I want to know that in a college whether male students prefer mathematics more than the female students or not or the male students prefer biology over the maths or not; in this case what we have to do we are simply going to count the number of male and number of female students who are preferring the subject.

So, here the numbers are going to be the observation on the discrete variable why it is called discrete? Because this number of students can only be an integer there can be 5 student and 6 student, but there cannot be 5.50 students. So, in this case we would try to see that what is the nature of association about the gender versus the subject.

Similarly, in case if some vaccine or some medicine is given to some patients, then we would try to see that how many patients are getting affected? If there are significant number of patients which are getting affected by the medicine then one can conclude that yes the effect of medicine and the number of patients they are associated. And we would like to see, what is the nature of association in this case of discrete variable?

So, whenever we have our discrete of or say counting variable, we would like to know whether male is to in preferred mathematics over female students or not. For example, and say in say another example we will consider we want to know if the vaccine given to the disease percent was effective or not right. So, these observation you see they are based on the counting of the two variables and in with discrete and in this case the variables are discrete in nature or and their values are obtained only as a number.

(Refer Slide Time: 08:13)



Association of Two Variables

Example: Ranked observations

- Two judges give ranks to a fashion model.
- Two persons give ranks to food prepared or their scores are ranked.

These observations are the ranks of two variables (two judges).

Similarly, there can be third situation for example, in a viva or enough fashion show the model or the candidate, they appear before that interviewers. For example, in the case of a fashion model for example, the model comes on the stage and there is a group of people who try to judge the performance and based on that they try to give the marks.

Now what do we expect? We expect that if a model is good, then all the judges will be giving the high risk course. And in case if the model is bad, then all the judges are going to give the lower risk course, but it is possible in real life that whenever a model come there will be certain number of judges who are giving higher score and certain number of judges who are giving lowest scores.

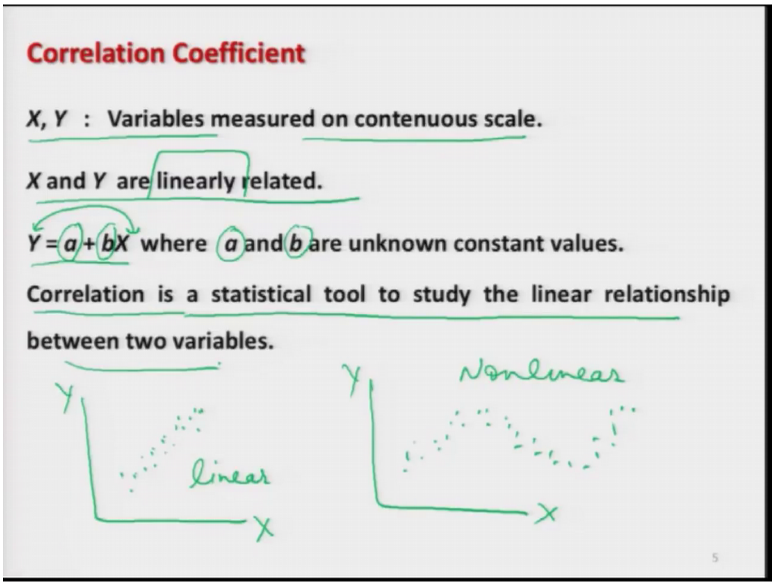
So; obviously, we would like to see that what is the correlation or what is the association or what is the nature of association between the ranks given by two judges how to obtain the ranks whatever the marks given by the judges, they are converted into ranks and finally, we would be interested in knowing the nature of association of those ranks not of their original values. So, in this case we have a concept of rank correlation coefficient. So, now we are going to consider here different types of thing.

So, in case of ranked observations there can be a situation for example, we had two judges give rank to the freshen model or there is another example that quite a person has cooked the food and there are two persons who are giving ranks to the food preferred or their scores are ranked. In these cases the observations are obtained as the ranks of two variables or say two judges or two persons are those two variables. So, now we have here different types of situations and those situations are described by the nature and behavior of the variable.

So, my objective is now here is that, I try to consider the nature of variable one at a time I try to show you that what are the different measures how to interpret them and how to compute them? So, in this lecture first I am going to consider that case where the variables are continuous and based on that we have a concept of correlation coefficient, after that when we are going for the ranked data then we will discuss about rank correlation coefficient.

And when we have counting variables then (Refer Time: 11:17) will be discussing about the different types of this coefficient like contingency coefficients chi square coefficient and so on. So, now we are going to start our discussion, where we have two continuous variable and we want to study the association between the two variables.

(Refer Slide Time: 11:41)



Correlation Coefficient

X, *Y* : Variables measured on contenuous scale.

X and *Y* are linearly related.

$Y = a + bX$ where a and b are unknown constant values.

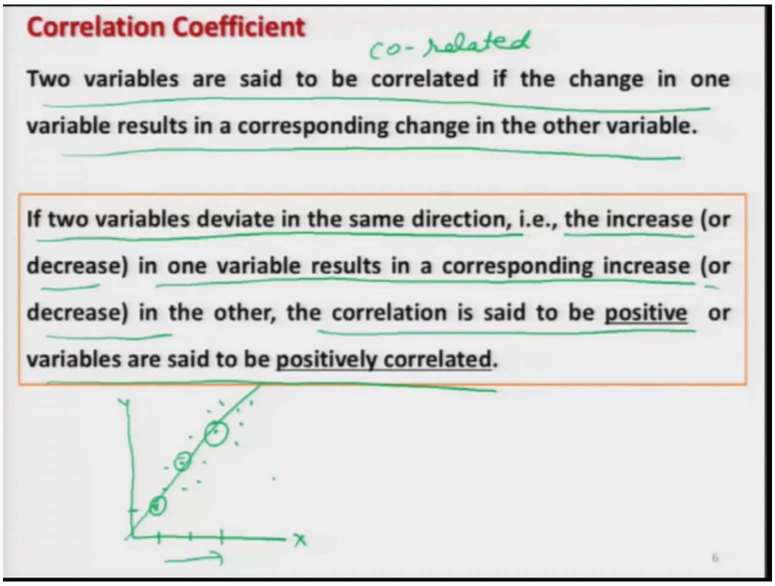
Correlation is a statistical tool to study the linear relationship between two variables.

What is the meaning of association? The association can be linear or that can be non-linear how do you know whether the relationship is linear or not? For example, we know and we have learnt that we will try to plot the data on with the scattered diagram. In case if the data looks to be like this then we say that there is going to be a linear trend and if the trend is like this means you can see here this will be called that the trend is not linear right ok.

So, now our basic framework is that we have here two variables capital X and capital Y and they are measured on a continuous scale. And both these variables are linearly related this is you have to keep in mind that now we are going to talk about the relationship which is linear in nature. So, we know that if the relationship is linear this can be express in a mathematical format by the equation of a line like as Y is equal to a plus b X; we have this a and b there are some unknown constant value. For example, in the case of line this equation is a and the standard form it is presented at y equal to mx plus c.

So, here this a is going to represent the c which is the intercept term and b is going to represent the value of m that is the slope of the line. So, now you can see here that this Y and X, they are related and they will have some degree of association. Now how to study this degree of association? For that we have a tool what is called as a correlation. So, correlation is a statical tool to study the linear relationship between two variables right.

(Refer Slide Time: 13:32)



Correlation Coefficient

co- related

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.

So, now means I can say that the two variables are said to be correlated if the change in the one variable results in a corresponding change in the other variable what does this mean? For example, we have taken the example of marks versus the time is spent in studies.

So, we know that when the students increase the time of their study then the marks obtained in the examination will also be changed. So, in this case the change in one variable is causing the change in another variable. Similarly in the case of say height and weight of small children, suppose height and weight are my two variables. So, when the height increases then usually the weight will also increases and similarly when the weight changes then the height also changes. So, change in weight also causes that change in height.

So, this is what we are trying to say and in these cases we can say that the two variables are correlated. If you try to see, how this word is coming correlated? This is co related. Now when we say that the two variables are correlated, so we are trying to say the change in the value of one variable is causing the change in the other variable.

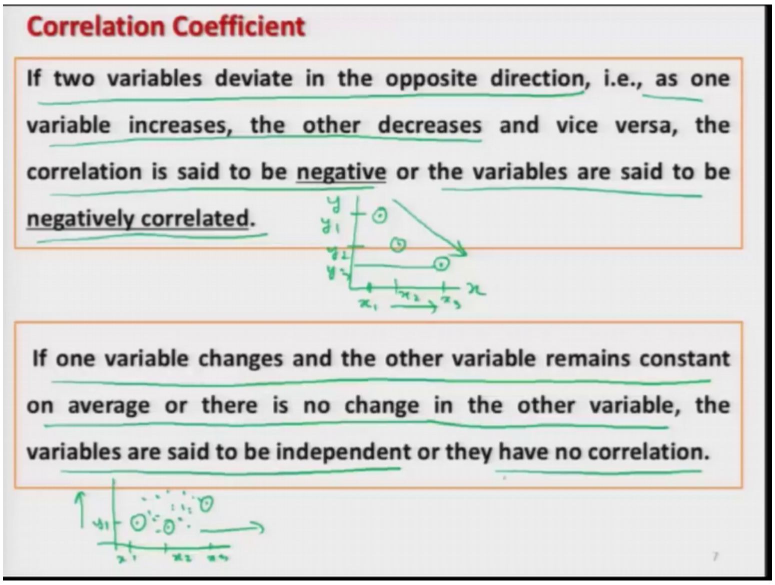
Now this change can be positive or this change can be negative means, if the change in one variable that is suppose if I try to change the value of one variable and suppose the other value that increases. Or in simple word if the values on X increase then the values of Y also increase; that means, it is a positive relationship and if the opposite happened that if the values of X increase, but the values of Y then decrease then this is negative relationship.

So, based on that we have a definition of positive correlation and negative correlation, so we can see here that if two variables deviate in the same direction; that is the increase or say equivalently the decrease in one variable results in a corresponding change to increase or to decrease the other variable. Then in this case the correlation is said to be positive or the variables are said to be positive correlated.

So, what will happen in this case? Suppose if I try to make here the scatter plot in the plot will if I try to change here the value of here X, suppose the Y value is here and if I try to increase the value of here X then y will also be somewhere here if I try to increase it more because will be here like this. So, we will have a graph like this.

So, in which the trend line will go like this. So in this case I can say that the observations on X and Y they are positively correlated and the nature of correlation is positive. So, in this case what is happening in case if the value of X is increasing then the value of Y is also increasing, the opposite is this if the values of X are increasing then the value of Y s are decreasing and the next situation is where in case if the values of x are increasing then something is happening in the y values, but the nature of the change is not clear.

(Refer Slide Time: 17:25)



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.

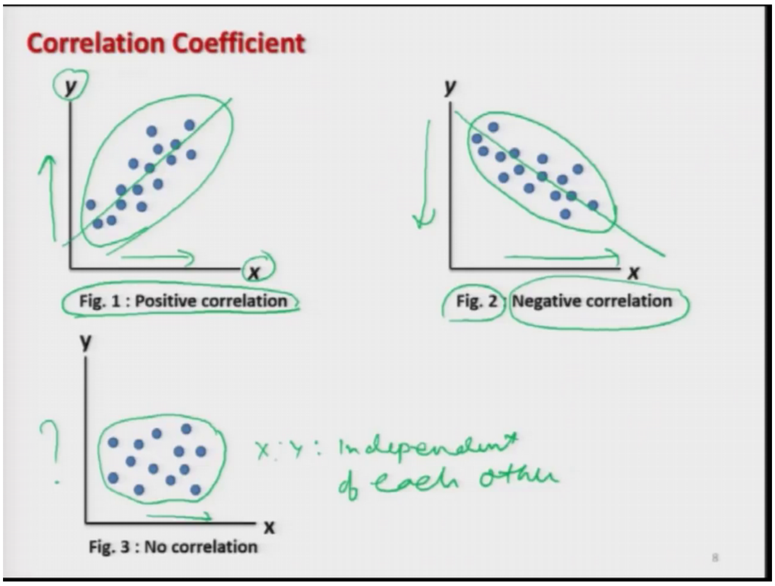
So, the next case is that if two variables deviate in the opposite direction what does this mean? That is as one variable increases then the other variable decreases or vice versa then in this case the correlation is said to be negative and the variables are said to be negatively correlated. So, in this case what will happen suppose if I try to plot the data, suppose I have value here of x as say x 1 and somewhere here is y equal to y 1 which is here like this right and then I try to increase the value of x say here x 2 then the value of y 2 becomes here which is lower than y 1. And similarly if I try to take here in the value of x 3 then this value comes over here which is here y 3.

So, in this case you can see here that as the values of x are increasing the values of y's are decreasing and there is a negative trend right. Similarly, if there are two variables and if the one variable changes and then the other variable remains constant on an average or there is very small change or no change in the other variable, then in this case the variables are said to be independent or they have no correlation.

For example, in case if I say I am trying to take here the value of say here x 1 and suppose this value comes out to be here y one somewhere here, then I try to take here x 2 and this value comes out to be here y 2 then in this case here x 3 and this value comes out to be here and so on we have some more values.

So, there is no clear cut trend in the data. So, this is indicating that when we are trying to change the value of x there is practically no change in the value of y, then in this case the variables are said to be independent of each other and they do not affect each other. And in this case we say that they have no correlation or they have 0 correlation.

(Refer Slide Time: 19:26)



Correlation Coefficient

Fig. 1 : Positive correlation

Fig. 2 : Negative correlation

Fig. 3 : No correlation

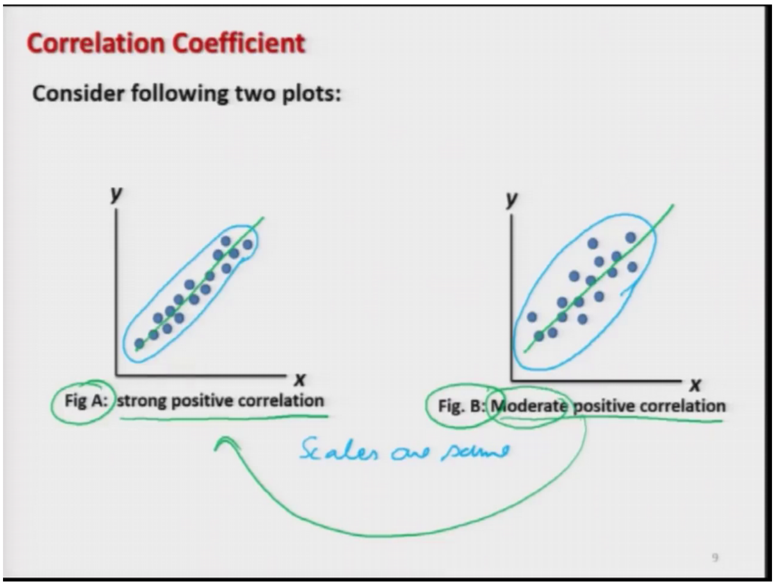
Now, in case if you try to represent these situations. So, what will happen here, suppose if I try to take the observations on two variable x and here y and we make a scatter plot then you can see here that in the figure number here 1 that these observations are here like this and in this case a trend line can be fitted which will passing through like this.

So, in this case when the values of x's are increasing then the values of y's are also increasing and so, we say that the relationship is positive and there is a positive correlation. Similarly in the figure 2 here as the values of x's are increasing then one can say here that the values of y's are decreasing and these values are these observations they are going like this and here in this case the trend line can be fitted like this.

So, in this case we would say that x and y are negatively correlated. And similarly if we try to change the value of x and suppose the values of x are increasing, but there is no change or say no trend in the values or we do not know how the y values are the y's are going to behave and they remain on an average as constant, there is no change when we are trying to change the value of here x.

So, it is here it is very difficult to say in this case we say that x and y they are independent of each other and in this case we say that the correlation between x and y is 0 or the x and y have no correlation here.

(Refer Slide Time: 21:07)



Correlation Coefficient

Consider following two plots:

Similarly, if I try to take here two situations of having the positive relationship, in say figure number A and figure number B, if you try to see the trend line will look like this. So, in figure number A you can see here that the points are lying more closer to the line then in the case of figure number B right.

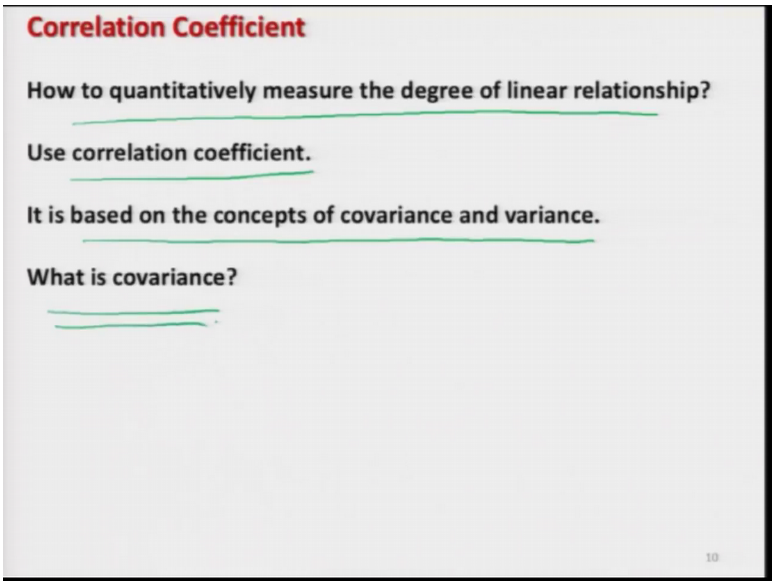
And means I am assuming that the scales are of both the figures are same means otherwise there can be a confusion. So, in this case when the points are lying close to the line then we say that, there is a strong relationship and in this case the relationship is positive. So, we say that there is a strong positive relationship between X and Y.

And similarly in the figure number B there is a positive relationship, but this relationship is not as strong as in the case of figure A. So, we call it that there is a positive coalition, but it is moderate. Now we have understood what is the concept of correlation? Now we need to define a quantity which can measure it and for this we have a definition of correlation coefficient and this correlation coefficient is based on the concept of variance and covariance what is variance that we already have discussed but what is covariance?

You can recall that in the case of variance what we had done we were trying to measure the variability of the observation around the mean value say arithmetic mean. Now if there are two variables and suppose both the variables are affecting each other they are interrelated, then when the value of one variable changes when the value of other variable also changes.

So, there is a sort of co variation between the two values. So, similar to the concept of variance we have a concept of covariance as variance measure the variability of a variable covariance measures the co variability of variables right.

(Refer Slide Time: 23:50)



Correlation Coefficient

How to quantitatively measure the degree of linear relationship?

Use correlation coefficient.

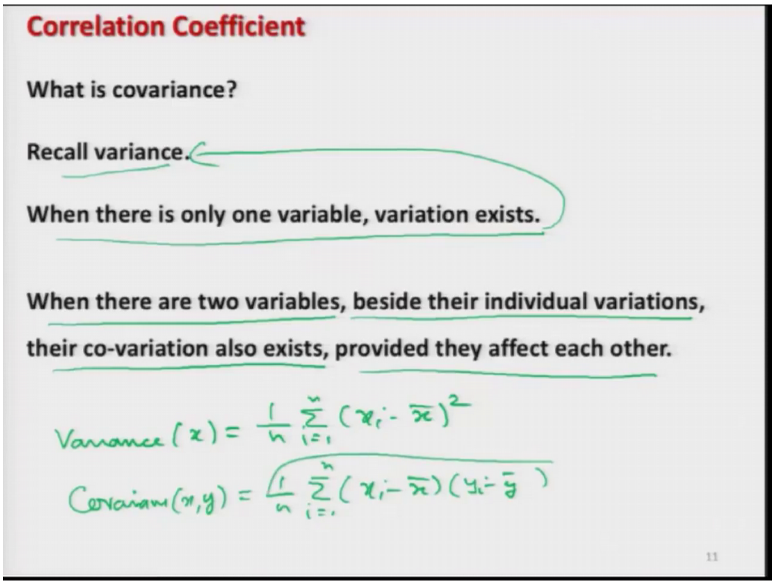
It is based on the concepts of covariance and variance.

What is covariance?

10

Now, we are going to address a question that how to quantitatively measure this degree of linear relationship and for that we are going to use the concept of correlation coefficient which is based on the concepts of covariance and variance and now first we address what is covariance?

(Refer Slide Time: 24:07)



Correlation Coefficient

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.

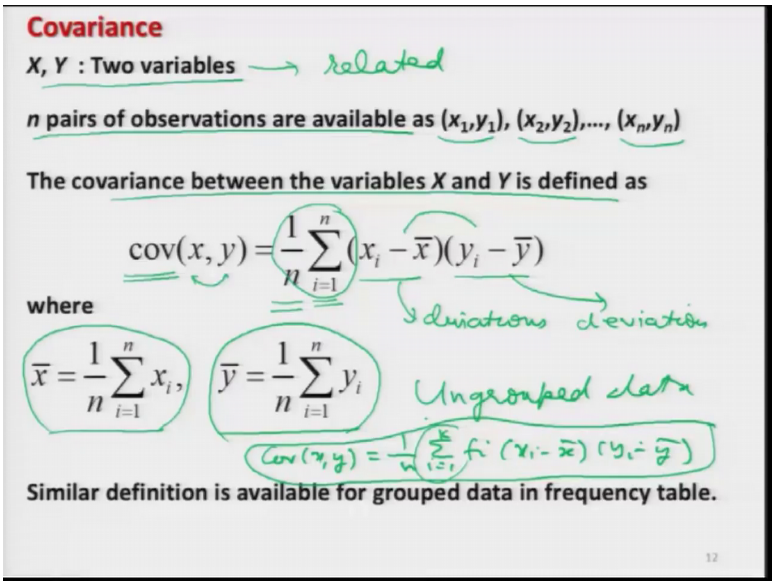
Variance $(x) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$

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

11

So, as we have discussed this covariance is very very similar to the concept of variance, when there is only one variable where it should exist and that is measured by variance; when there are two variables or even more than two variables then for these two variables their individual variation exist; that means, the suppose there are two variables of course, the two variables will have their own variance. And beside their own variants they will also have covariation and you have; obviously, if the effect each other if they are independent then there is no concept of co variation.

(Refer Slide Time: 24:59)



Covariance

X, *Y* : Two variables → related

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

$$\text{cov}(x,y) = \frac{1}{n}\sum_{i=1}^{n}\underbrace{(x_i - \bar{x})}_{\text{deviations}}\underbrace{(y_i - \bar{y})}_{\text{deviation}}$$

where

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

Ungrouped data

$\text{Cov}(x,y) = \frac{1}{n}\sum_{i=1}^{k} f_i (x_i - \bar{x})(y_i - \bar{y})$

Similar definition is available for grouped data in frequency table.

12

So, the question is now how to quantify it and how to measure it. So, but before going further you may recall that we had defined the variance of a variable here x say here 1 upon n summation I goes from 1 to n x i minus x bar whole square. Similarly, if what we try to do here, we try to write this function, as I go through 1 to here n and we try to write down here, x i minus x bar and say another x i minus x bar will be replaced by y i minus y bar.

And this will be sort of quantity that will be measuring the co variation between x and y. We assume that there are two variables which are represented as X and Y and it is obvious that we are assuming that these variables are related or correlated. Now we have obtained n pairs of observations on these two variable and these observations are expressed as x 1 y 1 x 2 y 2 x n y n.

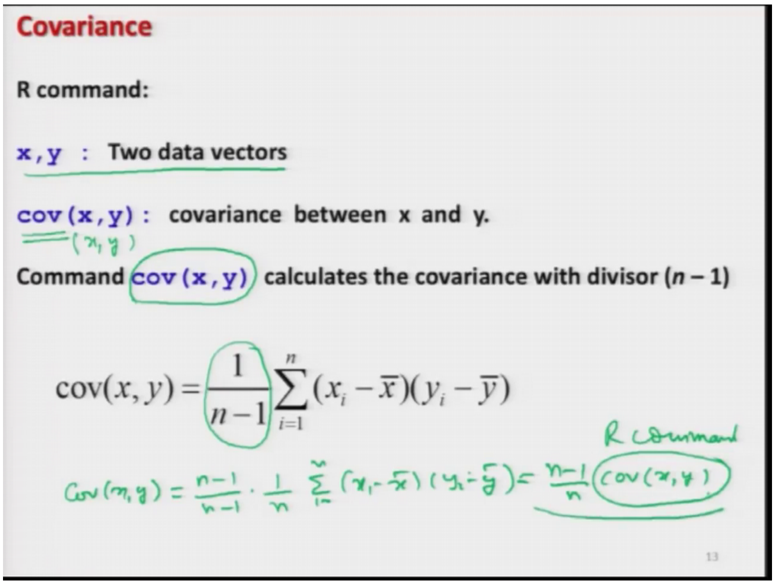
So, these are numerical values and we already have understood while discussing the graphical techniques that how to obtain such observations. Now the covariance between the variables x and y based on the sample observations this is defined as covariance cov indicating the covariance between two variables X and Y it is defined as 1 upon n summation I goes from 1 to n x i minus x bar y i minus y bar.

So, you can see here these are the deviations in x i and y i minus y bar they are the deviations in y i's. And we are trying to take a cross product of the deviations in x and deviations and y and we are trying to find out the average of those cross product of deviations. And here this x bar and y bar they are the sample means of x and sample means of y the sample mean of x is defined here like this and sample mean of y is defined here like this right .

And I have given you here the definition of covariance in case of actually in case of ungrouped data. And similarly in case if you want to have it for the group data then a then a similar definition can be defined as a covariance between x y is equal to 1 upon n I goes from 1 to here k see here fi x i minus x bar and y i minus y bar.

But you have to remember that here these symbols and notation they are going to have a different interpretation; we are now this x i and y is they are indicating the mid values and those x 1 x 2 x n y 1 y 2 y n that dictator had been grouped into k groups and so on, but anyway I will consider here only 1 ks.

(Refer Slide Time: 28:18)



Covariance

R command:

x, y : Two data vectors

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

(x, y)

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

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

R command

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

13

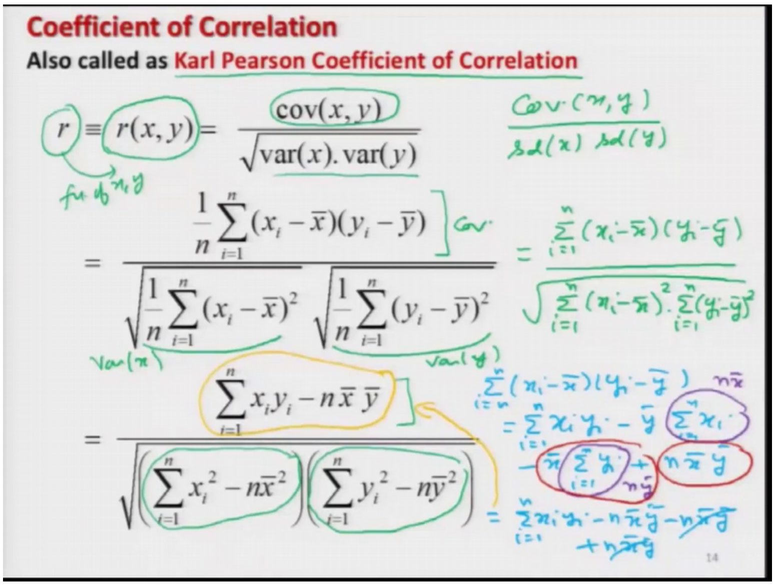
Now, the next question is how to compute this covariance on the basis of given set of data in the R software. So, if I try to indicate x and y are the two data vectors, then the syntax or the command to compute the covariance is a c o v all in a small letter and inside the arguments we give the data vector. But here you have to remember one thing that, this command c o v in the R software this will give us the value of covariance whether divisor here and minus 1.

So, in case if you want to find out the covariance between x and y say having a divisor n then what you need to do here that you need to multiply and divide by here n minus 1 into the quantity say here 1 upon n i goes from 1 to n x i minus x bar y i minus y bar and then this will be become here and minus 1 upon n and covariance between x and y and this is your here the r command right.

And this is the same story that we also had discussed in the case of variance that the variance was defined in two ways having a divisor n and say divisor n minus 1 and we had discussed that when we have the divisor n minus 1 then this is an unbiased estimator of the population mean. And the same story continues here also, that when we try to take the divisor to be n minus 1 then it is going to be an unbiased estimator of the population covariance, but anyway I am not going into the details of estimation and statistical inference.

So, but this is for your information. So, that in case if you really want to compute a particular type of covariance with divisor n or say n minus 1 you should at least know how to do it and you should also know what r is trying to give you ok. But anyway I will not take an example here to compute a to show you the on the r console, but that I will try to show you in the next lecture when I am trying to compute the correlation coefficient. Now I come to the definition of correlation coefficient.

(Refer Slide Time: 30:47)



Coefficient of Correlation

Also called as Karl Pearson Coefficient of Correlation

$$r \equiv r(x,y) = \frac{\operatorname{cov}(x,y)}{\sqrt{\operatorname{var}(x).\operatorname{var}(y)}} \qquad \frac{Cov(x,y)}{sd(x)\ sd(y)}$$

fn of x, y

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

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

$$\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) = \sum_{i=1}^{n}x_iy_i - \bar{y}\sum_{i=1}^{n}x_i - \bar{x}\sum_{i=1}^{n}y_i + n\bar{x}\bar{y}$$

$n\bar{x}$ $n\bar{y}$

$$= \sum_{i=1}^{n}x_iy_i - n\bar{x}\bar{y} - \cancel{n\bar{x}\bar{y}} + \cancel{n\bar{x}\bar{y}}$$

14

This coefficient of correlation is also called as Karl Pearson coefficient of correlation there is a reason, because Karl Pearson considered that the coefficient. So, if you try to see this correlation coefficient is denoted by r and this is equivalently written as r and inside the bracket xy. So, that this is indicating that this r is a function of two variables x and y and this quantity is defined as the ratio of covariance between x and y and square root of variance of x and variance of y. So, essentially if you try to see this is covariance between x and y divided by standard deviation or standard error of x and standard deviation or standard error of actually y.

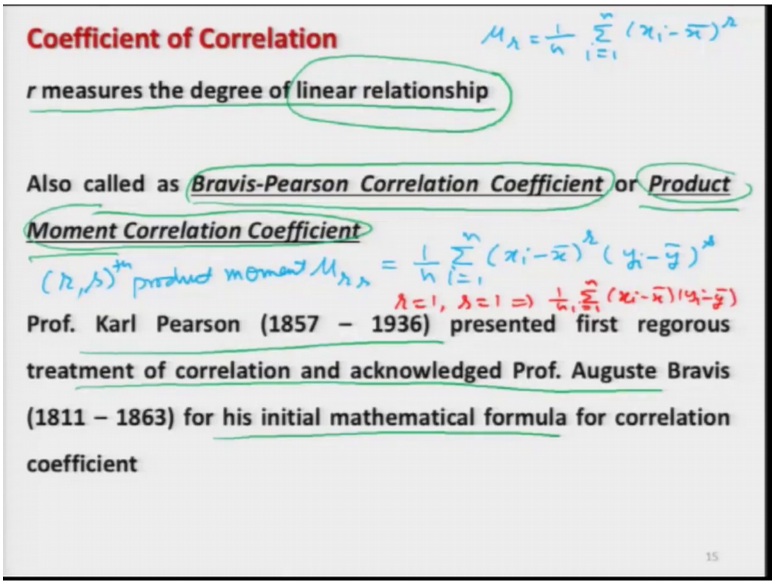
So, that is what I said that in order to define coefficient of correlation we need the concept of covariance and variance. So, we now know that this covariance is written like this one upon n summation x i minus x bar y i minus y bar this is covariance and this quantity here is trying to define the variance of x and this quantity here is trying to define the variance of y and then we try to take its square root. So, this is the expression for the correlation.

So, this can be simplified if you want to make it more clear this is summation x i minus x bar y i minus y bar divided by say summation i goes from 1 to n x i minus x bar whole square and summation, i goes from 1 to n y i minus y bar whole square right. And if you try to further simplify it then the numerator will become summation x i y i minus n times x bar y bar and this denominator summation x i minus x bar square will become summation i goes from 1 to n x i square minus n times x bar square and variance of y will become similarly summation y i square minus n times y bar square you may recall that we had solved this expression.

When we discussed the concept of this variability and while discussing the concept of variance. Similarly, if you want to see here that how this expression can be simplified which is involved in the definition of covariance summation x i minus x bar and y i minus y bar. So, this can be written as i goes from 1 to n x i y i minus see here y bar summation x i i goes from 1 to n minus x bar summation i goes from 1 to n y i plus n times x bar y bar. And now if you try to observe here what is happening what is this quantity?

This quantity is nothing but n times x bar and this quantity here is n times y bar. So, what will happen that the two factors here this and this will get canceled out and I can write down here that, here the summation x i y i i goes from 1 to n minus n times x bar y bar minus n times x bar y bar plus n times x bar y bar. So, this n this gets canceled out and we get here the same quantity what we have obtained here like this is the same quantity right.

(Refer Slide Time: 34:48)



Coefficient of Correlation

$\mu_r = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^r$

r measures the degree of linear relationship

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

$(r,s)^{th}$ product moment $\mu_{rs} = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^r(y_i-\bar{y})^s$

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

Prof. Karl Pearson (1857 – 1936) presented first regorous treatment of correlation and acknowledged Prof. Auguste Bravis (1811 – 1863) for his initial mathematical formula for correlation coefficient

15

So, having understood the basic definition or the mathematical form of the coefficient of correlation can us try to understand what it is doing and what are the different types of interpretation? So, essentially r is measuring the degree of linear relationship well, it is a very important term in a linear relationship well. You always have to keep in mind that correlation coefficient can be used only to measure the degree of linear relationship, in my experience I have seen that many people they try to use the correlation coefficient blindly and even they are trying to use this correlation coefficient to measure the degree of non-linear relationship, this is actually wrong.

And if you try to see this mathematical form of this correlation coefficient this is only a mathematical function, a mathematical formula whenever you try to give some value of x and some value of y it will give you some numerical value. But the interpretation of those values will be wrong and they will not be indicating the information contained inside the data.

So, this is my humble request to all of you that please use this correlation coefficient only to measure the degree of linear relationship and in order to do so first use the scatter plot try to see whether the relationship is linear or not that can be increasing or decreasing whatever you want. But the trend has to be linear and only then one should use the definition of correlation coefficient.

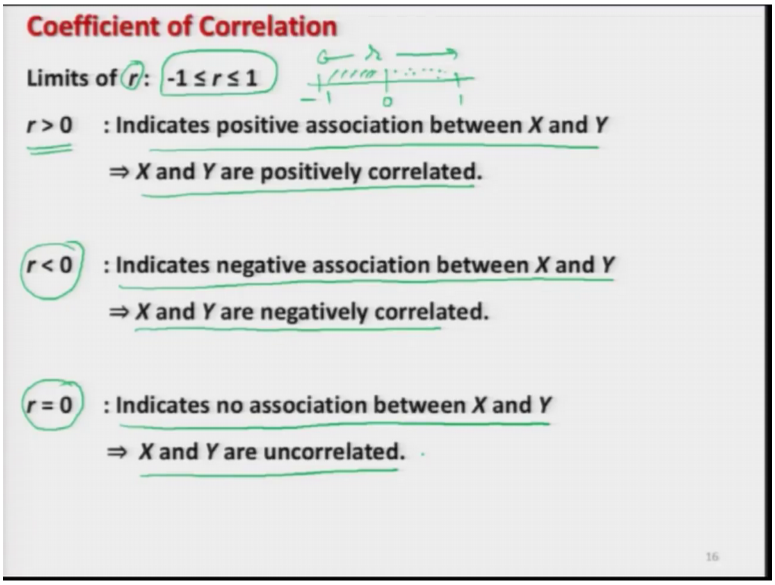
This correlation coefficient is also called as Bravis Pearson correlation coefficient and also as say product moment correlation coefficient why this is called as Bravais Pearson correlation coefficient? Actually Professor Karl Pearson presented the first rigorous proof or first mathematical rigorous treatment of the correlation and he acknowledged Professor Auguste Bravis because, Professor Bravis had made some initial mathematical contribution by giving the mathematical formula for the correlation coefficient.

So, that is the reason it is called as sometimes this is also called as Bravis Pearson correlation coefficient and this is also called as product moment correlation coefficient. why this is called as a predictive moment correlation coefficient? You might recall that, we had learned the definition of r h central movement and it was given as 1 upon n summation i goes from one to n x i minus x bar this power of here r.

So, this was a sort of the automatic mean of the rth power of the deviation in the value of x is. So, this was valid only when we have one variable, but now suppose if I have two variable what I can do here is the following, I can consider the deviations of x I can consider the deviations of y and then, I will try to take the rth power of divisions of xi s and s th power of the divisions of y i's then I will try to find out the automatic mean of the product of these deviations. And this is denoted as here mu r s and this is called as say r s th product movement.

So, that is why this is called the product moment correlation coefficient, why? Because in case if you try to substitute see here r equal to 1 and s equal to 1 this will give you the value here 1 upon n summation i goes from 1 to n x i minus x bar and y i minus y bar which is your covariance between x and y.

(Refer Slide Time: 39:12)



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.

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

$\Rightarrow$ X and Y are uncorrelated.

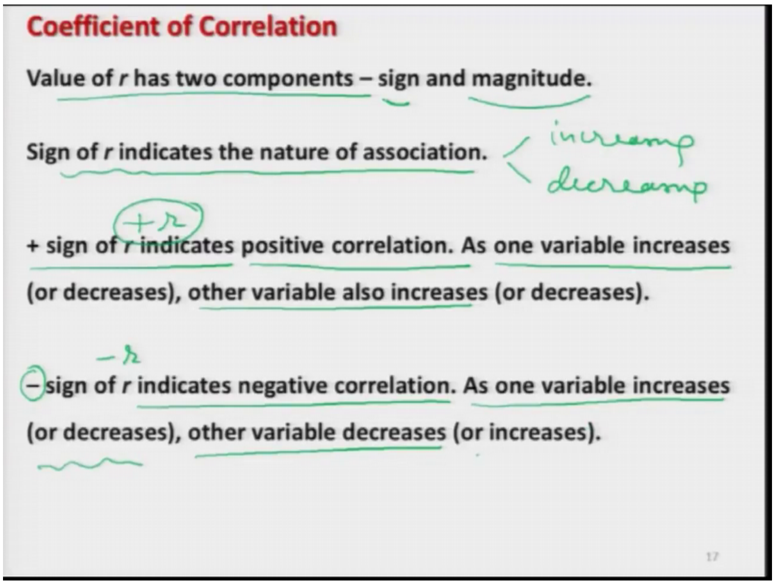
16

Now, we try to discuss the magnitude and sign of correlation coefficient what is their interpretation? So, this correlation coefficient value lies between minus 1 and plus 1 well, I am not giving you here the mathematical proof. So, r lies between minus 1 and plus 1 and those values 1 and minus 1 they are inclusive. So, what is the interpretation? You can see here this is here minus 1 and this is here 1 and this is here 0 so, this is the limit of here r.

So, what happens if r is negative lying in this side and what happens if r is positive lying in this side? So, when we try to compute the value of r on the basis or given set of data and if this comes out to be positive then, this indicates that there is a positive association between x and y and hence x and y are positively correlated. Similarly, if the value of r comes out to be negative, then this indicates that that there is a negative association between the two variables x and y and hence these two variable x and y are negatively correlated.

Similarly, if r equal to 0; that means, if you compute the value of correlation coefficient and it comes out to be 0 well, 0 is a critical value, but even if this is very close to 0, then in this case this indicates that there is no association between X and Y and hence X and Y are uncorrelated.

(Refer Slide Time: 40:54)



Coefficient of Correlation

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

Sign of *r* indicates the nature of association. increasing decreasing

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

+r

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

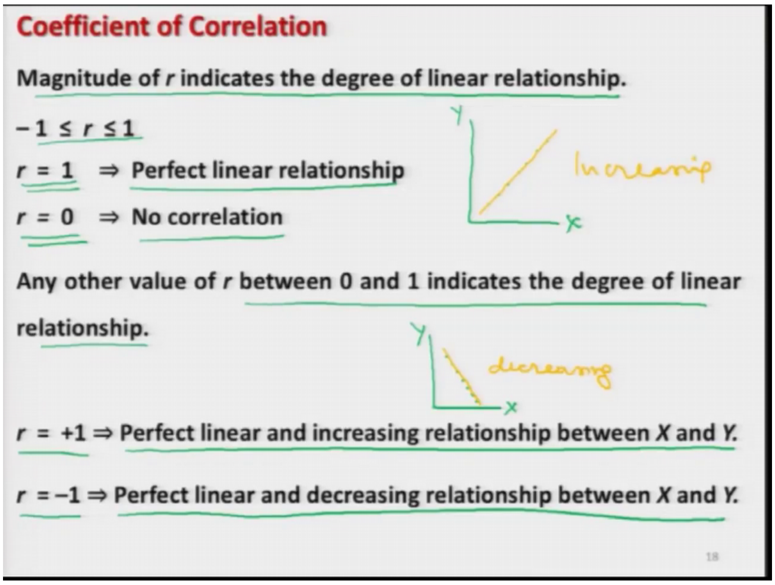
–r

17

So, now we have seen that the value of r has two components; one is the sign and another is the magnitude. This sign of correlation coefficient indicates the nature of association; that means, whether the relationship has got an increasing trend or decreasing trend right. So, the positive sign of r indicates that there is a positive correlation, this means what as one variable increases or the values of one variable increases then the value of other variable also increases. And similarly if the values of one variable decrease and the values of other variable also decrease.

So, plus r will give us an information that the relationship is positive and the degree of linear relationship is the magnitude of r. And similarly if we consider the negative sign of r say minus r, then the negative sign indicates the negative correlation. So, as the values of one of the variable increases then the value of other variable decreases so the relationship is opposite and similarly, if the value of one variable decreases then the value of other variable increases.

(Refer Slide Time: 42:33)



Y
X
Increasing
Y
X
decreasing

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*.

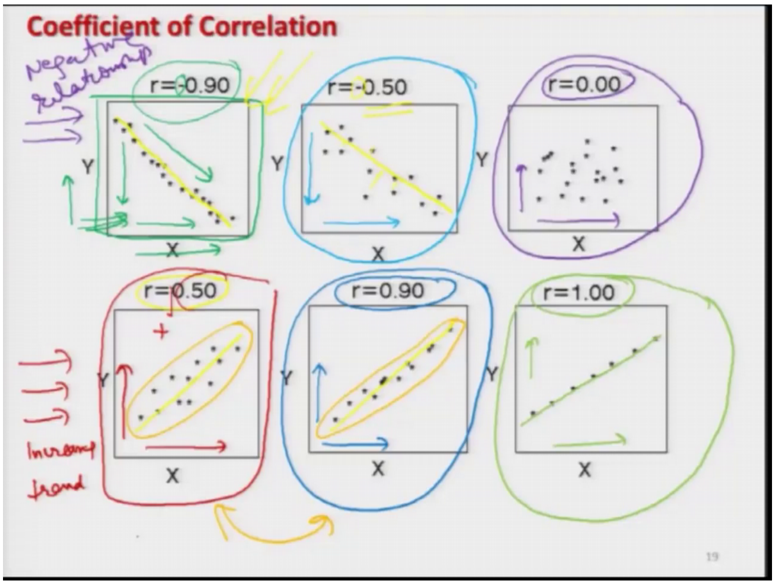
18

And what about the magnitude of r the magnitude of r indicates the degree of linear relationship so, we have seen that our lies between minus 1 and plus 1. So, there will be two extremes are the minus 1 and plus 1 and one is the middle value. So, when we say r is equal to 1, this indicates the perfect linear relationship what does this mean that if I try to plot the scatter plot between x and y then all the points are lying exactly on the same line.

So, if we try to make here a line on this graphic it will look like this. So, there is a 100 percent perfect relationship and all the values are lying exactly on the line. Similarly, if I say r equal to 0, this will indicate that there is no correlation there is 0 correlation and in case if I say any other value of r between 0 and 1, that will indicate the degree of linear correlation relationship higher the value of r higher the degree of linear relationship this relationship can be positive or negative.

So, when r is equal to plus 1 this will indicate the perfect linear and increasing relationship between x and y and when we say r equal to minus 1 then this will indicate a perfect linear and decreasing relationship between X and Y. For example, if you try to make it here x and here y then what will happen this in case of decreasing relationship this relationship will be like this and in this case if you try to make a trend line this will be a perfect straight line. So, this is the case of decreasing relationship and in the above one was the case of increasing relationship right.

(Refer Slide Time: 44:32)



Coefficient of Correlation

And now I would simply try to show you these things graphically what does this actually mean for example, now what you have to do, you simply have to keep your concentration on my pen right. First you try to look over here at this figure, you can see here that this X is here X is indicating the values of X and this is here the values of Y inside all these figures.

So, we can see here in this picture here that as the values of X are increasing the values of Y's are decreasing here like this. And in case if you try to create here that line trend line this trend line will be something like this and you can see here that all the points are not lying exactly on the same line.

So, in this case the sign of the correlation coefficient will be negative and definitely this value is not 1 so, but most of the points are lying very close to the line so the correlation coefficient can be close to 0.90. And similarly in the next figure in this one you can see here when we are trying to increase the value of X the values of Y's are decreasing and in this case if you try to create the trend line this will be like this. So, now in case if you try to compare this picture then most of the points are lying close to the line, but these points are not as close as in the figure number 1 here.

So, that is why if you try to see the values of here r this is minus 0.50 this is indicating that the relationship is linear by this negative sign and the value here is 0.50 well its not 0.50, but it is very close to half and this is lower than the value of 0.90 as in the earlier case. Now in the third case, you can see here that there is no clear relationship and the value of X's are increases Y's are also changing and there is no relationship in this case. So, this is the case of 0 correlation or no correlation and this is represented here by r equal to 0.00. So, all these two cases in the above panel, they are trying to indicate the negative relationship.

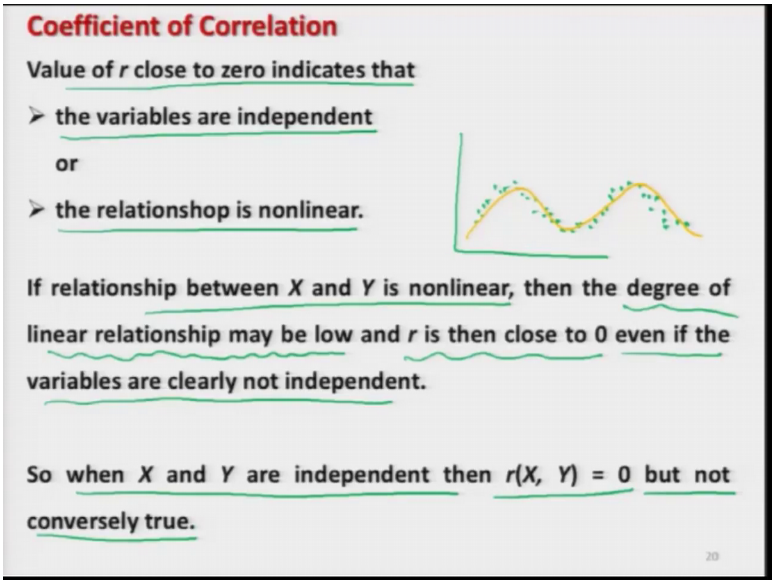
Now, we try to consider the lower panel, where we have the increasing trend here you can see here in all the panels there is an increasing trend in the data. So, if you try to see here in the first picture here the as the values of X's are increasing the values of Y's are increasing. And if you try to make here a trend line you can see here it is like this, but definitely the points are not so close to the line so the value of r is plus 0.50 here.

So, that is indicating that the sign here is positive. So, that is indicating the increasing relationship and 0.50 is the magnitude of the r which is indicating that well there is a linear relationship, but; obviously, all the points are not lying exactly on the line. And similarly, if I try to increase the value of correlation coefficient something like here r is equal to 0.90 the sign is positive in this picture and you can see here as the values of X are increasing, the values of Y's are also increasing and the trend line here will be like this and if you try to compare the first two pictures over here this picture and this picture, you can see here that in this reason the points are lying closer to the line in comparison to the points in the first picture like this.

So, that is why this difference is indicated by the magnitude of the correlation coefficient which is from 0.50 to 0.90. Finally, in the last picture here you can see that as the values of X are increasing Y's are also increasing and all the points are lying exactly on the same line. So, this is the case of perfect increasing linear relationship and this is indicated by the value of r is equal to plus 1.

So, this is how we try to get the information about the magnitude and direction of the relationship by looking at the scatter diagram. So, there are 6 diagrams I have represented here and they will give you a fairly good idea that how the things are going to be done. Now I would try to address a very important and a very interesting observation.

(Refer Slide Time: 49:40)



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.

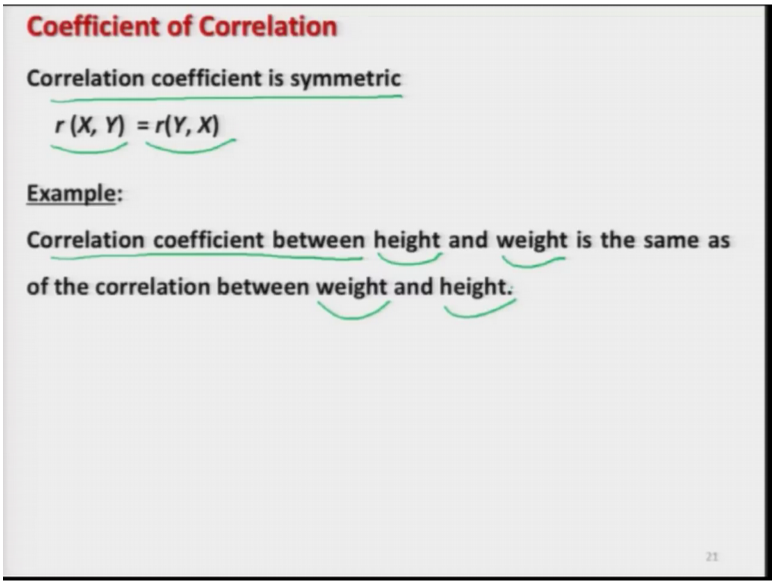
20

You see whenever the value of r is close to 0 or say r equal to 0, this may indicate two things, there can be two types of interpretations either the variables are independent or the relationship is non-linear why because correlation and coefficient is only indicating the degree of relationship when it is linear.

So, now what happens if the relationship between X and Y is non-linear, in this case the degree of linear relationship computed by the correlation coefficient may be low and so the value of r is close to 0 and this will be indicating that as if the variables are independent, but it is not correct because their existed a non-linear relationship. So, in this case this r is close to 0 even if the variables are clearly not independent for example, if I say there is a trend like this one you can see here this that the relationship is very very clear there is a sort of sine curve, but the correlation coefficient in this case will give you the value close to 0.

So, be careful with these types of interpretation and remember that when X and Y are independent then the correlation coefficient between X and Y will be equal to 0, but not the opposite, but not the converse is true. So, this is very important point for you to keep in mind that what is the meaning of r equal to 0?

(Refer Slide Time: 51:24)



Coefficient of Correlation

Correlation coefficient is symmetric

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

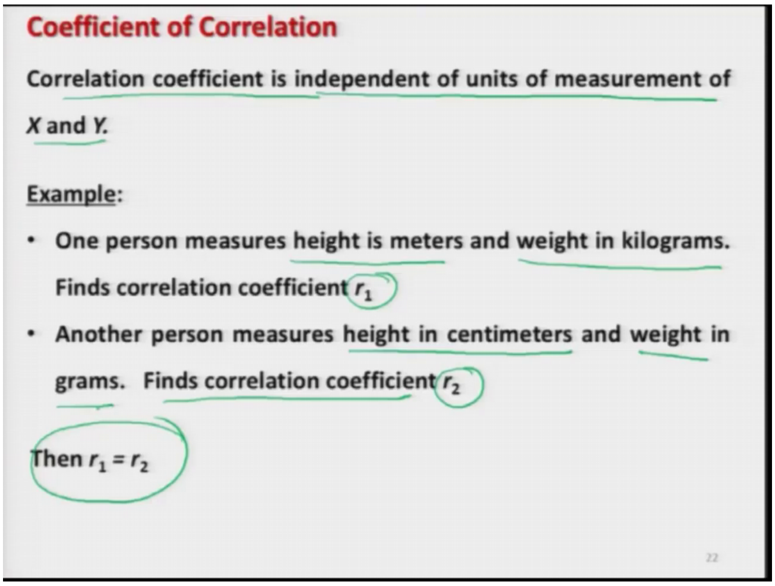
Example:

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

21

Similarly, another property correlation coefficient is symmetric; that means, correlation coefficient between X and Y is the same as correlation coefficient between Y and X what does this mean? That if somebody finds the correlation coefficient between height and weight and say and other person find the correlation coefficient between weight and height then both are going to be the same.

(Refer Slide Time: 51:49)



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

Similarly, one very nice property with correlation coefficient has that this quantity is independent of the units of measurement in X and Y. So, what is the advantage? Suppose one person measures the height in meters and weight in kilograms and find out the correlation coefficient say r 1. Now there is another person who measures the height and weight of the same set of people, but he measures the height in centimeters and weight in grams and he finds the correlation coefficient as say I 2, then in this case both r 1 and r 2 are going to be the same they will have the identical value and that is a very nice advantage of using the correlation coefficient.

Now in this lecture I would stop here that was a pretty long lecture. And my objective was to give you the information and development of the correlation coefficient concept. Now, in the next lecture I will show you that how to compute this on the R software and how to interpret it. In the meantime you please try to read from other books and try to develop the concept of correlation coefficient in more depth and I will see you in the next lecture till then good bye.