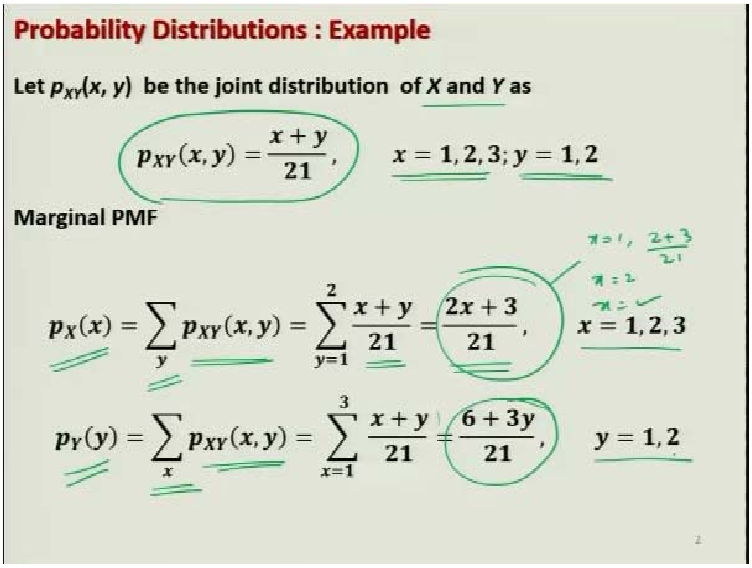
Essentials of Data Science with R Software- 1
Professor Shalabh
Department of Mathematics and Statistics
Indian Institute of Technology Kanpur
Lecture 48
Covariance and Correlation – Examples and R Software

Hello friends, welcome to the course Essential of Data Science with R Software- 1, in which we are trying to understand the basic concepts of probability theory and statistical inference. Well, you can recall that in the last lecture we had a good discussion on covariance and correlation, and we have understood the meaning, utility, and interpretation. Now, in this lecture I will continue with the same thing, covariance and correlation but I will try to give you some examples which are based on theory and I will try to show you that how you can compute this covariance and correlation in the R software. So, let us begin our lecture and try to see how we can do it.

(Refer Slide Time: 1:02)



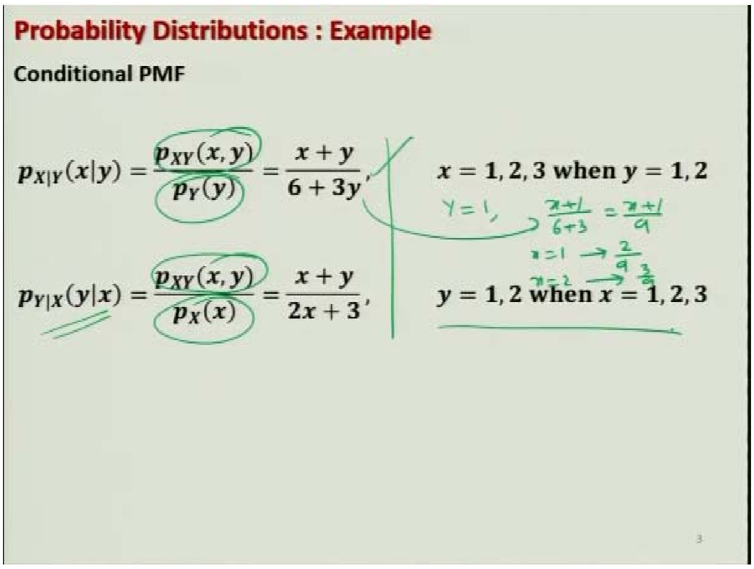
So, the first example I am going to take this is about the discrete random variable. Now you know computing covariance and correlation they are simply the function of joint probability and marginal distributions, once you can obtain those probability distributions either joint or marginal, you can compute their expectation, you can compute their variance and from there you can just compute.

So, whenever you are trying to compute the correlation coefficient, the calculation may be little bit longer because you have to find couple of things but lengthy does not mean difficult, the calculations will be simple. So, now we consider here a probability mass function for the joint random variables X and Y, so X and Y are two discrete random variable which have got this following a joint probability mass function which is given by $\frac{x+y}{21}, \; x = 1,2,3; y = 1,2$.

Well that is a very simple example to show you how you can proceed further, now from their first we need to find out the marginal probability mass function, so we find out the p(x) which is the marginal probability mass function for x which is obtained by summing over the values of y on this joint probability mass function, you simply substitute the value obtain the expression this will come out to be $\frac{2x+3}{21}, \; x = 1,2,3$.

So, for the given value of x you can compute this px. For example, for x equal to 1 this probability mass function will be 2 plus 3 upon 21 and so on. And similarly, you can compute for x equal to 2 into x equal to 3 also. Similarly, if you try to find out the marginal probability mass function for the y, then this is obtained by summing over the values of x over the joint probability mass function and if you can obtain this summation from here $\sum_{x=1}^{3} \frac{x+y}{21}$ you can obtain here this value, and this is true for y equal to 1 and 2.

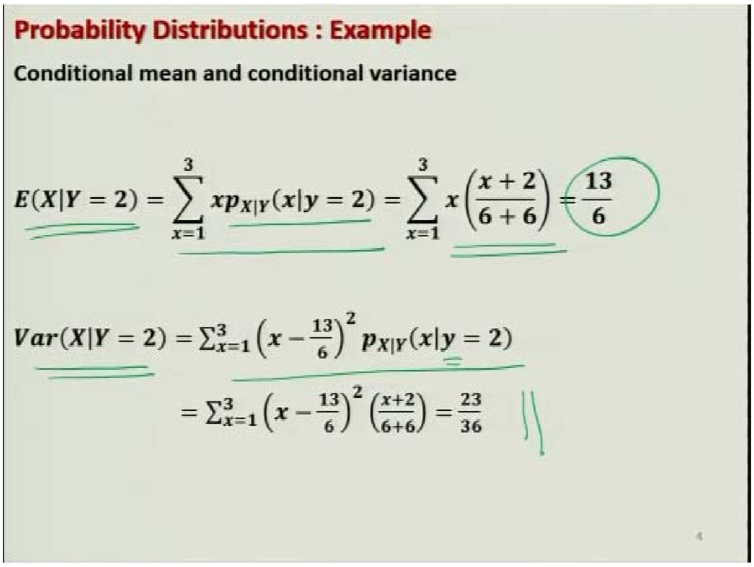
(Refer Slide Time: 3:12)



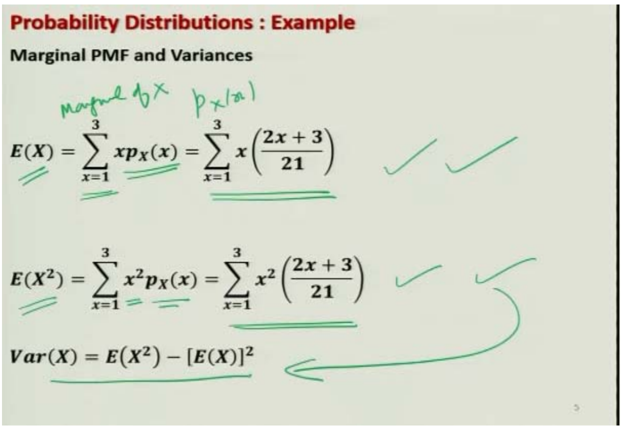
And now from here we can also find the conditional probability mass function that you can see just by taking the joint probability mass function divided by the marginal of y and similarly the join the conditional probability mass function of y given x, the joint probability mass function divided by marginal of x and these values can be obtained here like this without any problem.

And so for example, if you want to compute the conditional probabilities from this probability mass function, so here the y is given for example, if I take here y equal to here 1 that mean then this probability mass function is going to be (x +1)/6 +3 so this will become here x plus 1 upon 9 and now for x equal to 1 this value is going to be 2 upon 9, for x equal to 2 this value is going to be 3 upon 9 and so on. So, this is how we try to interpret these conditional probabilities and similarly you can do for conditional probability of y given x also.

(Refer Slide Time: 4:14)

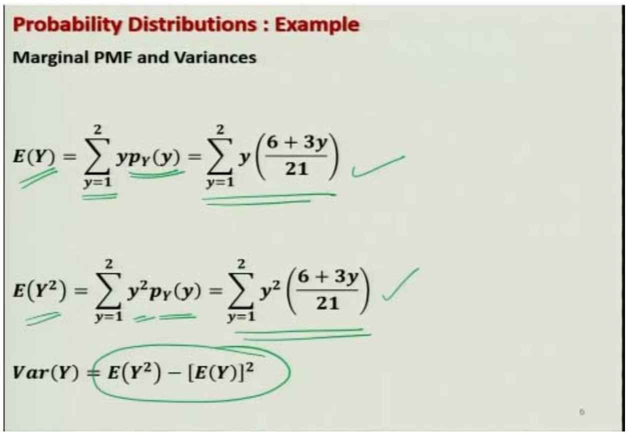


So, now we try to find out the conditional mean and conditional variance. So, from the f(X|y) we can find out here the E(X|y=2), so you can see here for example, this value is going to be here like this, and you simply have to substitute here y equal to 2 and then you can obtain here its value 13 by 6. And similarly, if you want to know the Var(X|y=2), you simply have to take the expression and then simply you have to substitute here y equal to 2 and just solve this value, it is as simple as that. (Refer Slide Time: 4:52)



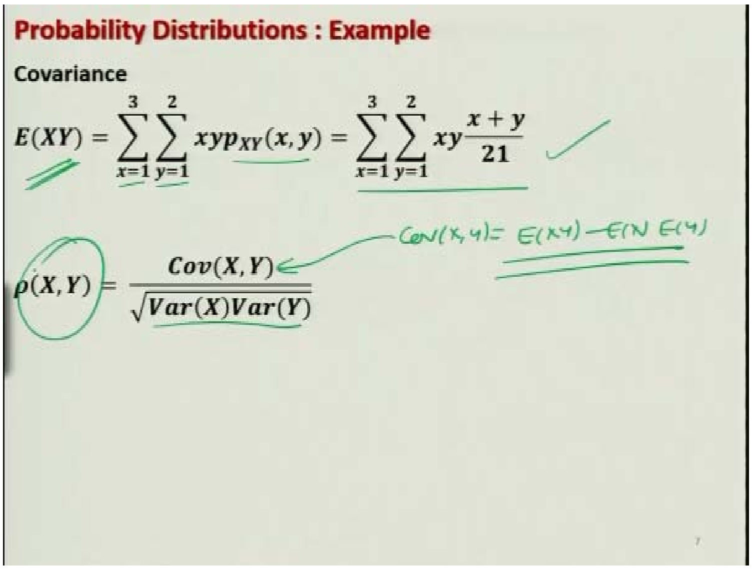
I am just trying to give you the basic idea and then from the marginal probability mass functions you can compute the simple expected value of x and variance of x also for example, if you want to compute the expected value of X first you need to find out the marginal of x that is px(x), so that you already have found so you can simply take here the values of $\sum_{x=1}^{3} x p_X(x)$ which is here like this and whatever is the value you can solve it and similarly if you want to compute the $E(X^2)$ you can simply obtain here the $\sum_{x=1}^{3} x^2 \left(\frac{2x+3}{21}\right)$ and this will be value here and you can simply solve it and just using this value and this value here in the expression for variance you can very easily compute the variance of x, that is not difficult.

(Refer Slide Time: 5:44)



Same thing you can do with the y also, that is you can use the marginal probability mass function of Y and then using this thing you can compute the value of your expected value of y where the summation is going from y goes from 1 to 2 and then you can simplify this value. Similarly, if you want to find out the expected value of Y square just try to take here summation y goes from 1 to 2 into the probability mass function of y this is the marginal distribution and then you get here this expression just substitute this marginal probability mass functions and then solve it you will get this value and you substitute these values over here in this expression of variance of Y, you can very easily compute the variance. My objective of showing this example only to show you that how are you going to write out these limits whether you have to write down the limit of x or y and so on.

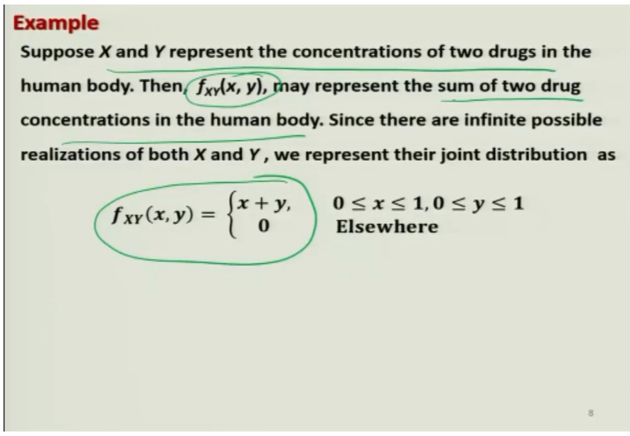
(Refer Slide Time: 6:40)



Now, similarly you can compute here the $E(XY) - \sum_{x=1}^{3}\sum_{y=1}^{2} xyp_{XY}(x,y)$ and you can simply solve whatever is the value you get here you just obtain it and then try to compute and now here covariance between X and Y that you can obtain as a E(XY) – E(X)x E(Y) and you already have obtained all these expressions.

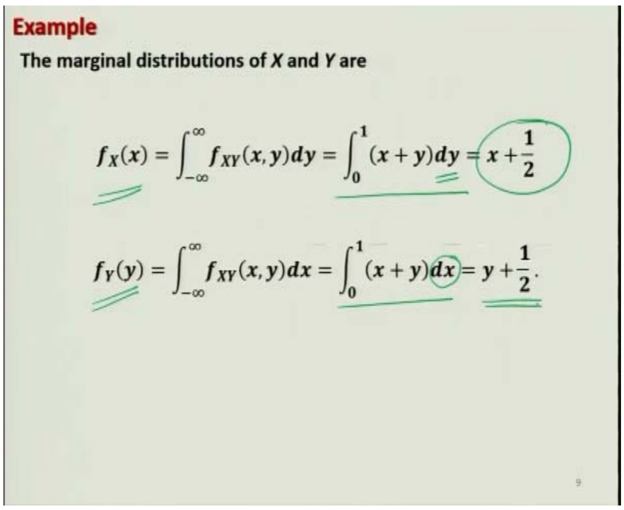
So, you can compute this covariance of XY, substitute it here you already have computed the values of variance of X and variance of Y you compute them, and you can very easily compute the correlation coefficient. Well, I am not interested in the value, but I want to show you the methodology how are you going to do it.

(Refer Slide Time: 7:29)



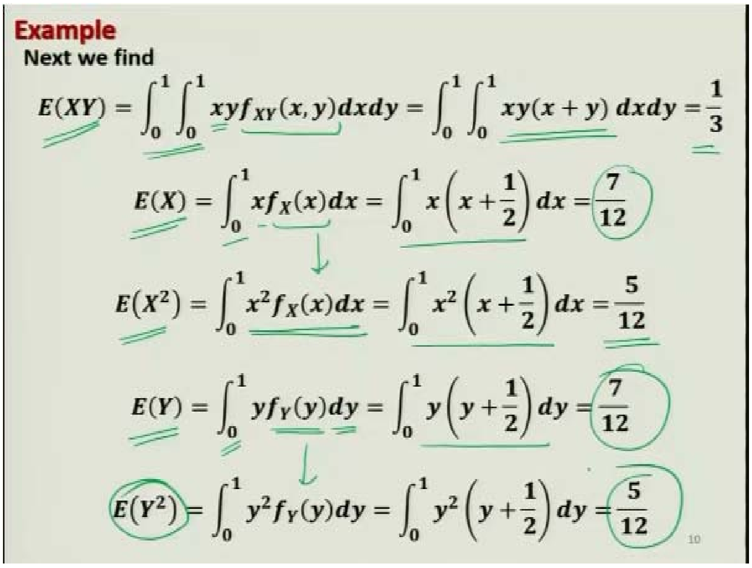
Similarly, I can do the similar excise with the continuous random variable case. Suppose X and Y are the two continuous random variables, and this is the same example actually that we considered earlier also where we have a joint PDF of x and y, where x and y are representing the concentration of the two drugs the human body and this fxy is representing the sum of the two drug concentrations in the human body. So, now this joint ODF is given by this thing $f_{XY}(x,y) = \begin{cases} x+y, & 0 \le x \le 1,\ 0 \le y \le 1 \\ 0 & \text{Elsewhere} \end{cases}$.

(Refer Slide Time: 8:07)



You know from this PDF we try to find out the correlation coefficient but for that we have to first compute a couple of quantities. So, first we try to find out the marginal distribution of X and Y that we already had obtained earlier but just for the sake of review this f(x) is obtained $f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x,y)dy = \int_0^1 (x+y)dy = x + \frac{1}{2}$.

(Refer Slide Time: 8:49)

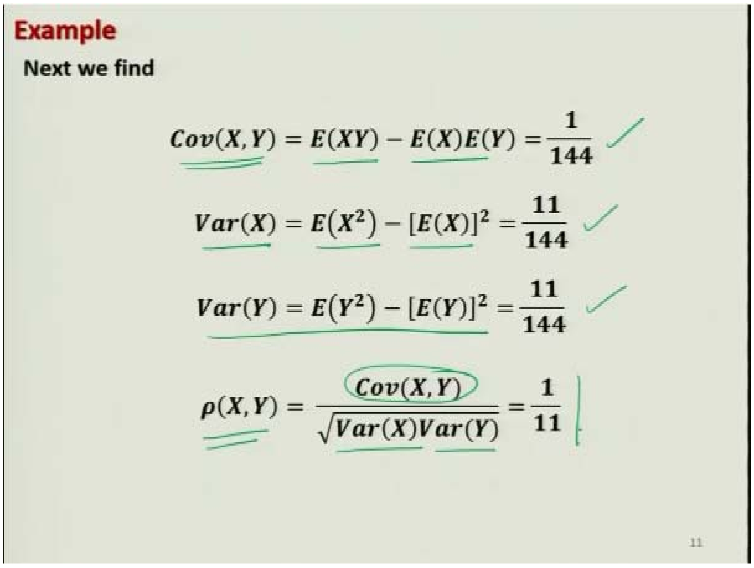


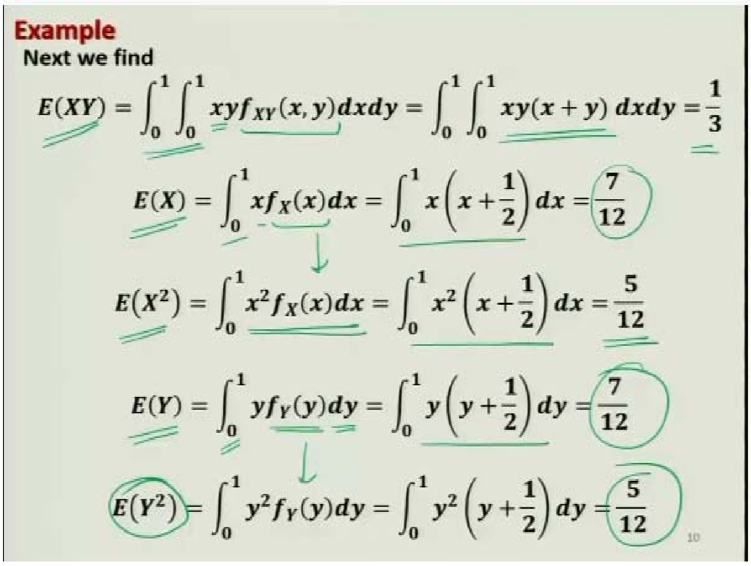
And then we try to obtain different types of expectation they are very simple using this joint and marginal PDF. So, if you try to find out the $E(XY) = \int_0^1 \int_0^1 xy f_{XY}(x,y) dxdy$. So, if you try to simply write down these expressions and try to solve it you will get here the value 1 by 3.

And similarly, if you want to find out the expected value of X for that you have to use the marginal distribution of x, so that will be 0 to 1 x into marginal of x which you already have found, so you simply try to substitute these values and solve it you will get here 7 by 12 and similarly if you try to find out expected value of X square exactly on the same line as you have done here you can just write here $\int_0^1 x^2 f_X(x) dx$ and then you can write down all these values and simplify it you will get here the value 5 by 12.

And similarly you can find out $E(Y) = \int_0^1 y f_Y(y) dy$ and if you substitute all these values you will get here the value 7 by 12 and similarly you can compute expected value of Y square exactly in the same way and you will get here the value 5 by 12, no issues I am 100 percent confident that you can solve these integrals,.

(Refer Slide Time: 10:08)



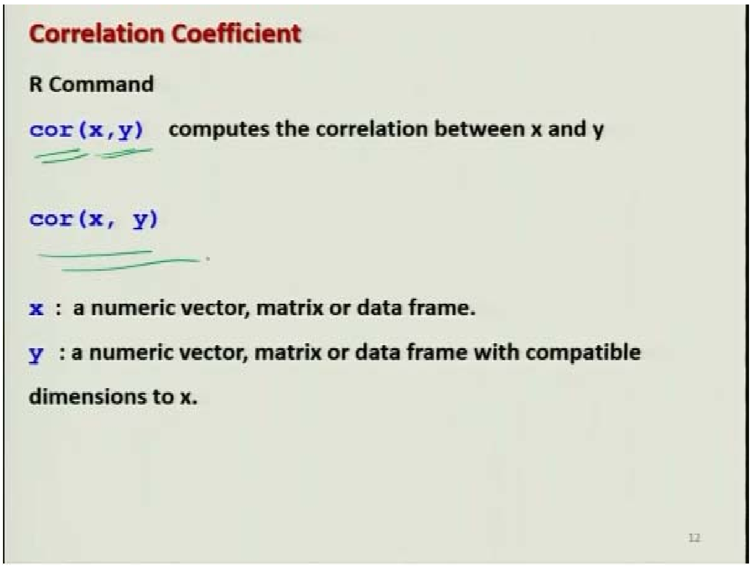


Now using this expression you can find out the value of covariance as E(XY) – E(X)E(Y), you already have obtained here all these values, all these numerical values, just substitute you will get here this value 1 upon 144, you can obtain the here the value of X as $E(Y) = \int_0^1 yf_Y(y)dy$and then $Var(Y) = E(Y^2) - [E(Y)]^2 = \frac{11}{144}$

, and just substituting all these values in the expression for the correlation coefficient you can find out that expression $\rho(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}} = \frac{1}{11}$, so you can see here this is

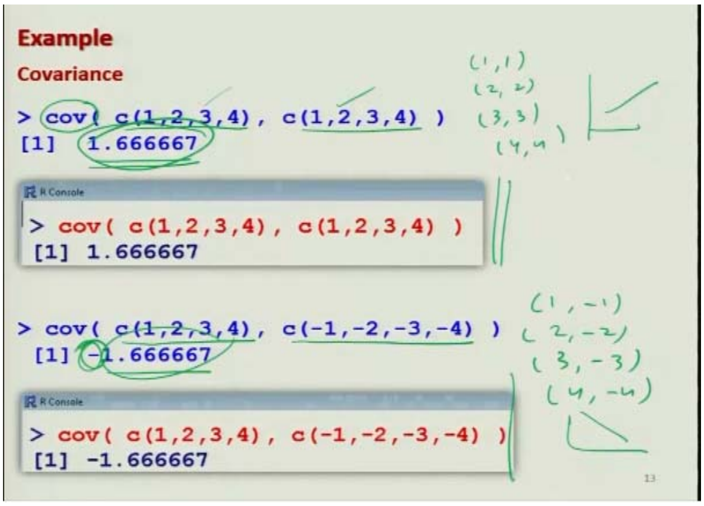
how you can theoretically compute the correlation coefficient for any given joint probability density function.

(Refer Slide Time: 11:02)



Now, in case if you want to compute this quantity in the R software then the command here is cor, and then inside the parenthesis you have to write down here the data vectors x and y, so they are going to need to indicate the paired observation in which the x and y observations are stored in two different data vectors and this cor x, y will give you the value of correlation coefficient.

(Refer Slide Time: 11:33)

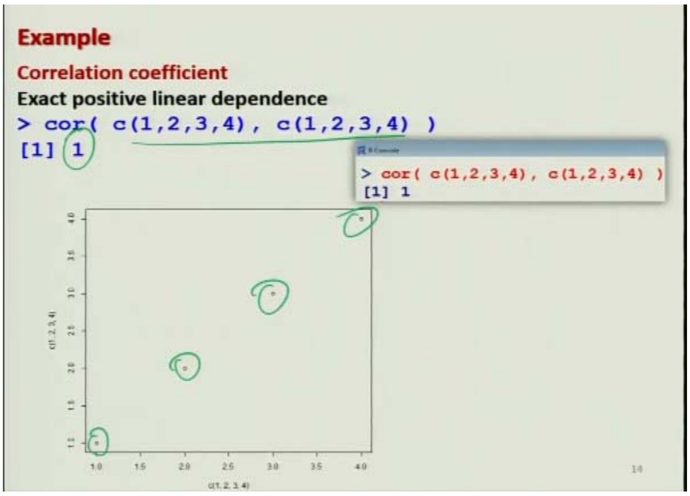


Well, if you go to the help menu of this cor function you will see that there are different other types of correlation like as rank correlation etcetera, but here definitely we are not considering them, if you wish you can learn them yourself but we are thinking or we are talking here of only Karl Pearson correlation coefficient which is measuring the degree of the linear relationship between two quantitative variables.

Now, I try to show you something, that if you try to take here only two data vectors 1, 2, 3, 4 and 1, 2, 3, 4 and you try to find out the covariance, you can see here the value is coming out to be 1.66 and this is here the screenshot and if you try to take first data vector in which all the values are positive and the second data vector in which all the values are negative -1, -2, -3, -4 you can see here the magnitude remains the same but the sign here is now negative.

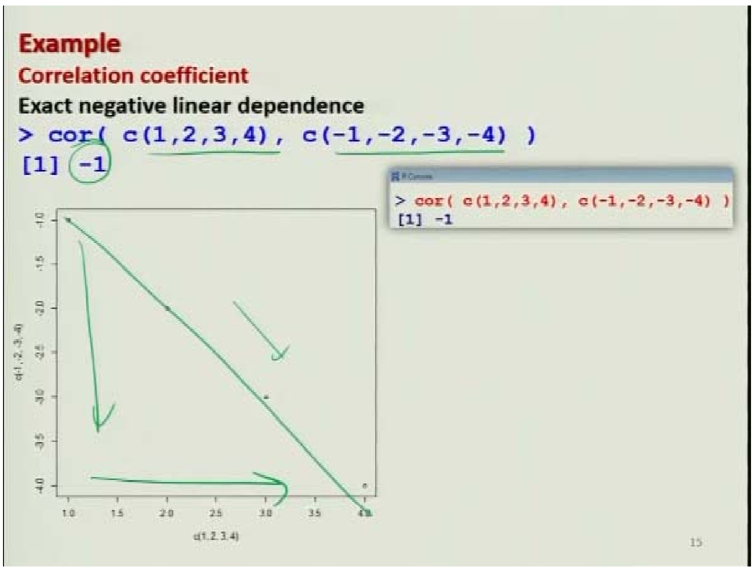
So, this is what I meant, that if values of x and y, actually the values are here like this 1, 1, 2, 2, 3, 3 and here 4, 4. So, if you try to see the magnitude here one here and here they remain the same at 1.66 but their directions are opposite plus and minus, so this is what I told you that the sign of the covariance indicates whether x and y are increasing or decreasing with respect to each other. So, in this case you can see here the points are 1 -1 and 2 -2, 3 -3 and here 4 -4 like this so you can see here the relationship is increasing, whereas here the relationship is decreasing and this is being indicated by this -sign.

(Refer Slide Time: 13:22)



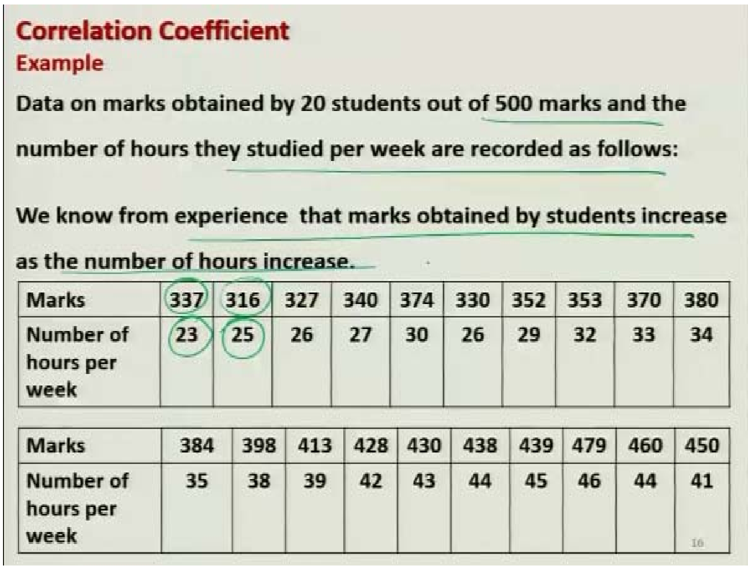
And this is here the screenshot I will try to show you on the R console and in case if you try to plot them here they will look like this first point, second point, third point, fourth point and for this value 1, 2, 3, 4 if you try to find out the correlation coefficient so that is the correlation coefficient between actually to exactly the same data sets. So, this will come out to be 1 because that is the case of perfect linear relationship.

(Refer Slide Time: 13:51)



And similarly, if you try to take here one data vector to be 1, 2, 3, 4 and another to be -1, -2, -3, -4 that means the directions are opposite, so you can see here that these points are lying on this line which is decreasing and this correlation coefficient between these two data vectors comes out to be -1, so you can see here that the trend here is like this that as the value of x are increasing the values of y are decreasing, so this is what I meant that when you are trying to find out this perfect negative linear relationship.

(Refer Slide Time: 14:23)



Correlation Coefficient

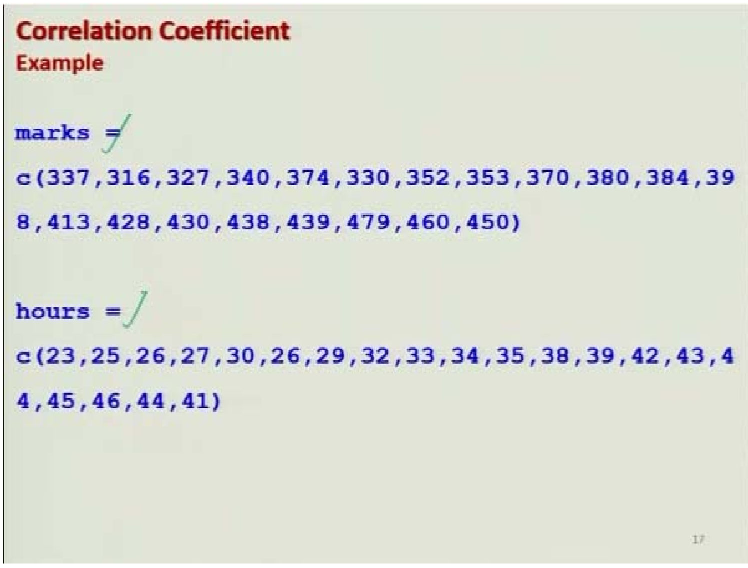
Example

Data on marks obtained by 20 students out of 500 marks and the number of hours they studied per week are recorded as follows:

We know from experience that marks obtained by students increase as the number of hours increase.

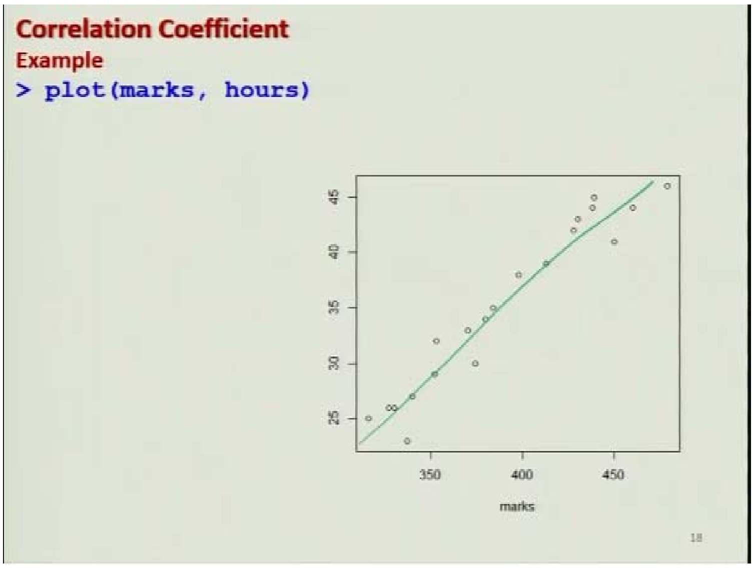
Marks	337	316	327	340	374	330	352	353	370	380
Number of hours per week	23	25	26	27	30	26	29	32	33	34

Marks	384	398	413	428	430	438	439	479	460	450
Number of hours per week	35	38	39	42	43	44	45	46	44	41



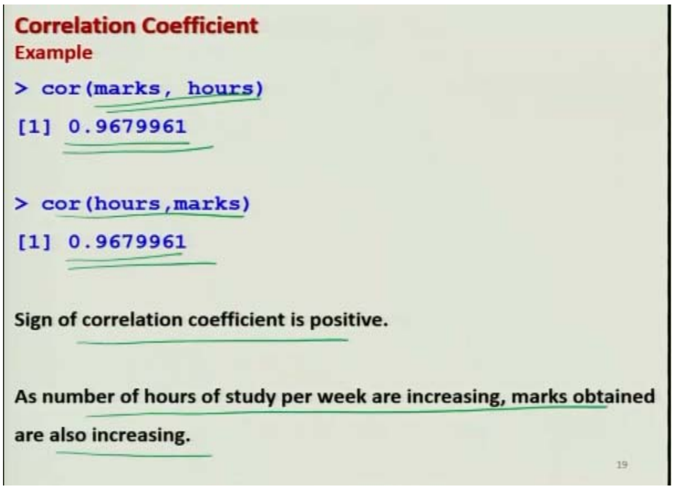
Now, I try to take a very simple example and try to show you that how these things are going to be measured. Suppose we have 20 students and we have obtained their marks out of 500 and then we also have asked those students that how many hours in a week they have studied, and their data is obtained like this that the student number 1 that student has studied 23 hours in a week and that student has got 337 mark. Similarly, the student has studied 25 hours in a week and the student has got 316 marks and so on. So, we know from our experience that the marks obtained by the student the increase as the number of hours increase.

(Refer Slide Time: 15:09)



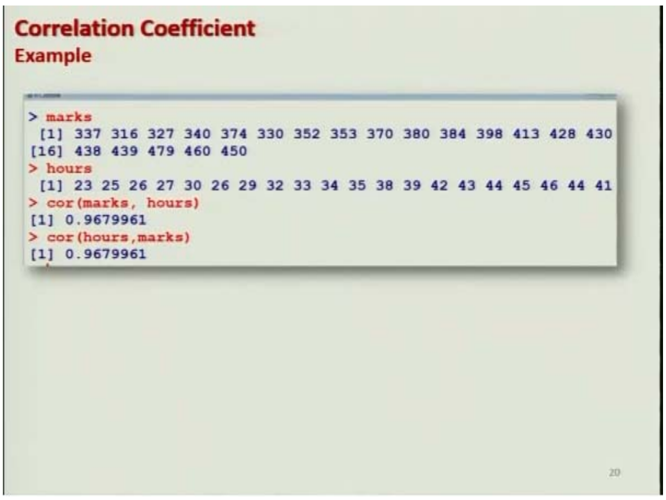
So, now let us try to understand this analysis through the R software. so I try to give this the marks in two data vector marks and hours and now I try to plot, so you can see here all these points are lying nearly in a straight line and by looking at this you can say that lines are not very far away from the line, so that is giving us a that the degree of linear relationship here is quite high, that means points are or these two variables marks and hours they are strongly positively correlated. So, this is the graphical view that how you can look about the value of the correlation coefficient.

(Refer Slide Time: 15:53)



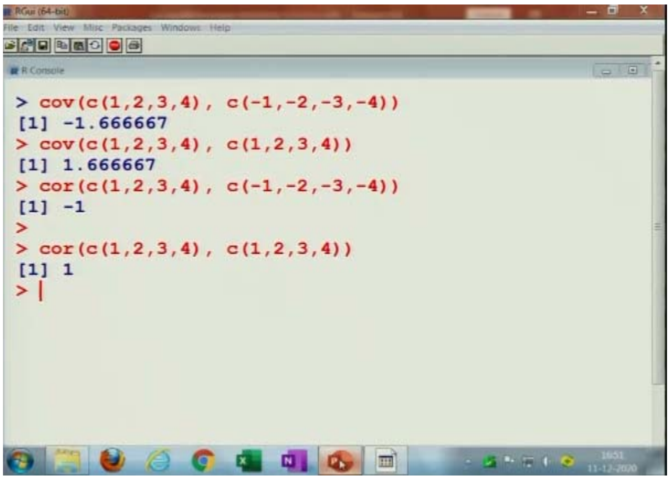
And if you try to find out the correlation coefficient between the marks and hours you can see that this value is coming out to be 0.96 which is very high mid close to 1. So that is what is being indicated by this graph also and if you try to reverse the roles of variable, that means you try to find out correlation between hours on marks you can see that this is the same because we know that correlation between x and y is the same as correlation between y and x and here you can see that the sign of the correlation is positive, so we can conclude now that as the number of hours of studies are increasing the marks obtained by the students are also increasing and that is according to our experience also this is what we believe.

(Refer Slide Time: 16:39)



And this is here the screenshot of the same operation.

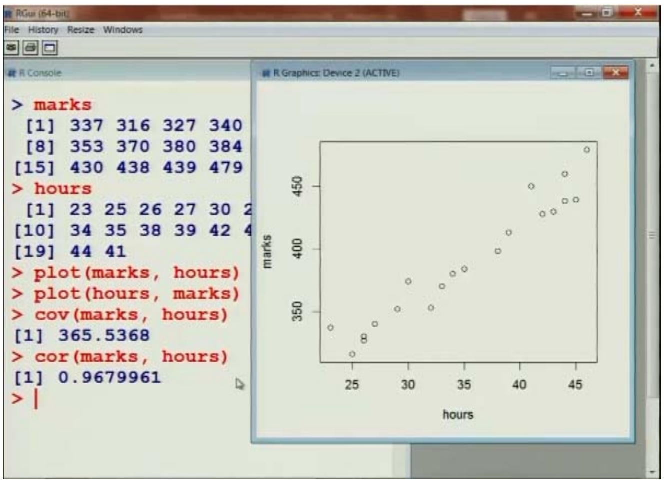
(Refer Slide Time: 16:51)



Now, I try to give you these things on the R console so that you can understand what is really happening here. So, if I try to find out first here is covariance between c(1, 2, 3, 4) and between c(-1, -2, -3 and -4). You can see here this value comes out to be like this -1.66 but if you try to remove these signs and you make both the observations to be positive, you can see here the sign becomes here negative, now the sign becomes here positive. So, this is how the direction of the or the kind of the covariance indicates that what is really happening, whether the trend is increasing or decreasing.

Now, in case if I try to find out the correlation coefficient between 1, 2, 3, 4 and -1 -2, -3, -4 then it is coming out to be -1 and if you try to find out the covariance between the same set of observation 1, 2, 3, 4, 1, 2, 3, 4 you can see here that this is coming out to be plus 1. So, these are the cases of perfect linear positive and negative relationship.

(Refer Slide Time: 18:16)



Now, I try to take this example and I try to enter these marks to save some time and then we try to do the further analysis. So, you can see here that these are here the marks, here hours and you if you try to create here a plot between marks and hours you can see here like this, that means ideally you can see that there is a straight line, and if you means it try to change this x and y axis that makes a better sense also, means number of marks depends on the number of hours also, even then you will see that the curve is remain the same, the trend remain the same, only the values are shifting, that does not make any difference.

So, now in case if you try to find out here the covariance between here marks and hours you can see here this is coming out to be like this and so that means it is positive. Now, in case if you try to find out the correlation coefficient between this marks and hours you can see that this is coming out to be 0.96. So, you can see here that this value is indicating the same what the information you are going to get from the curve well.

So, this is all about the covariance and correlation coefficients and now we come to an end to this lecture and you can see here this was a very simple and small lecture and my objective was to give you an idea that how you can calculate the covariance correlation etcetera, based on the probability mass function and probability density functions and I also wanted to show you that how you can compute these values in the R software which are very simple and straight forward.

But the main thing is this how to get the correct interpretation of those things and one thing I would say now we are venturing towards the outcome from the software also you can see that it also requires some practice and experience before you can get or you can understand what the data is trying to tell you in the correct way. So, if you practice more, you will become a better data scientist that is what I can say, so you practice take a data set try to plot the curve, try to find out the covariance correlation and I will see you in the next lecture till then goodbye.