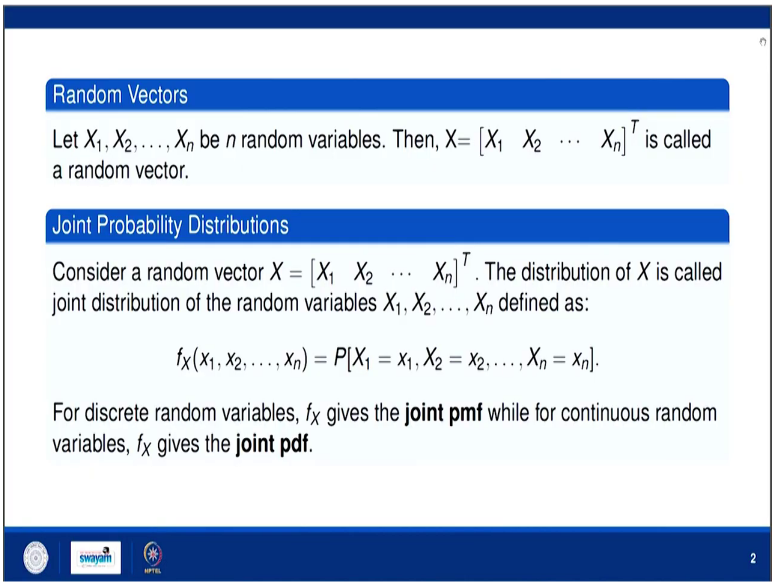
Essential Mathematics for Machine Learning
Prof. S. K. Gupta
Department of Mathematics
Indian Institute of Technology, Roorkee

Lecture - 35
Joint Probability Distribution and Covariance

Hello friends, welcome to lecture series on Essential Mathematics for Machine Learning. In the last two lectures, we have seen some basic concepts of probability we have seen that how we can find conditional probability, base theorem, distribution functions, their probability mass function, density function, etcetera. Now, in this lecture we see we will see that how we can find Joint Probability Distribution and in and using this how we can find Covariance.

(Refer Slide Time: 00:54)



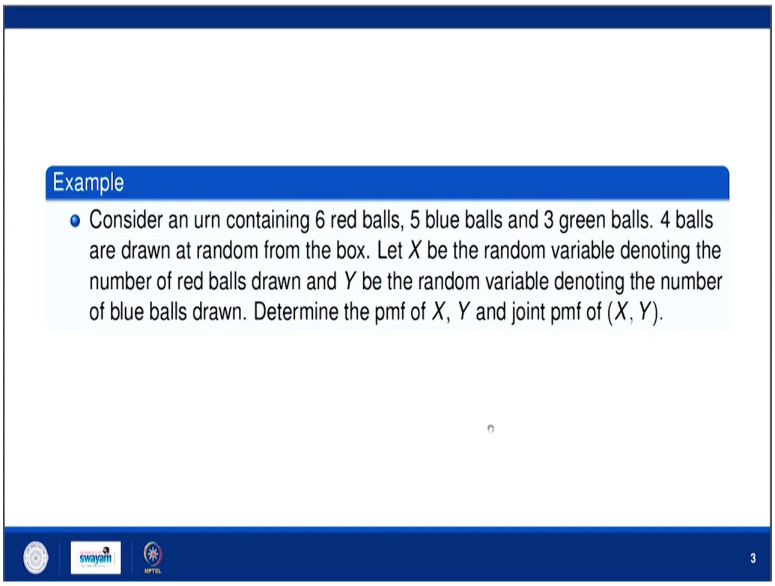
So, let us start now we are having random vectors, what random vectors are; so let X 1, X 2 up to X n be n random variables, so these are n random variables. Then a vector X which is simply X 1, X 2 up to X n this transpose, this is basically called random vector that means, if the collection of all the n random variables that is called random vector.

Now, joint probability distribution now consider a random vector X which is X 1, X 2 up to X n whole transpose, the distribution of this X; this random vector X is called joint distribution of the random variables X 1, X 2 up to X n defined as f x of x 1, x 2 up to x n. This means that probability when capital X 1 is taking x 1 value, capital X 2 is taking x 2 value, and capital X n is taking X n value. So, this is basically called joint distribution of random variables X 1, X 2 up to X n.

Of course, for discrete case if random variables are discrete, then this f x is simply we call it, joint probability mass function joint pmf ok. While for continuous case, if x 1, x 2 up to x n are these are continuous variables, continuous random variables; then we this is called this f x is called joint probability density function ok.

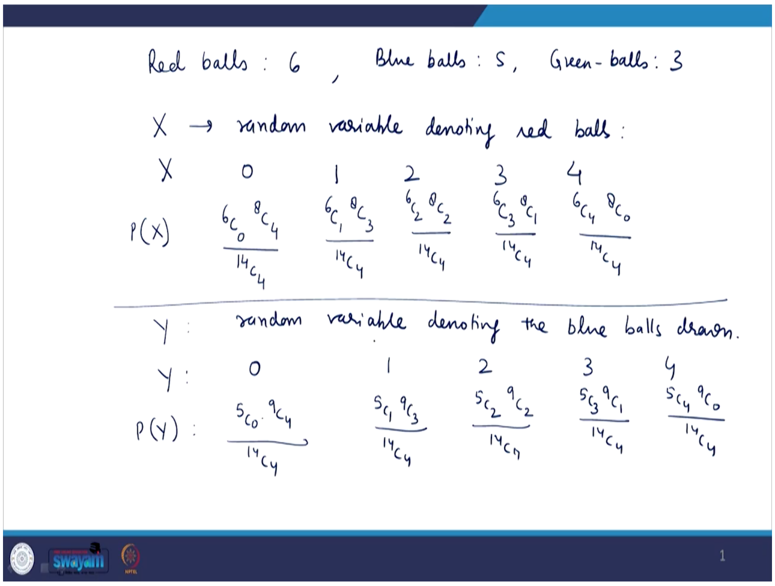
So, basically till now we have seen that how what is probability mass function or probability density function for a single random vector x, but now instead of single random x we are having n number of random variables. So, now instead of finding for a single random variable we are finding for random vector. So, now instead of calling it pmf or pdf we are calling it joint pmf or joint pdf, this is the only difference.

(Refer Slide Time: 03:02)



So, let us try to understand the concept of joint pmf or pdf using few examples. So, consider this example for joint pmf – joint probability mass function. So, what this example is now consider an urn containing 6 red balls, 5 blue balls and 3 green balls.

(Refer Slide Time: 03:25)



So, what we are having, we are having how many red balls? So red balls are 6 ok, then we are having blue balls 5, 5 blue balls, and then we are having green balls are 3 as per the question. Now, 4 balls are drawn at random from the box, the box containing total how many balls? It is 5 plus 6 plus 5 – 11; 11 plus 3 – 14 balls; total 14 balls, we are taking out any 4 balls.

Now, let X be the random variable denoting the number of red balls drawn and Y be the random variable denoting the number of blue balls drawn. Then we have to find out probability mass functions of capital X and capital Y and also a joint probability mass function of X, Y; so how can you find, let us discuss.

So, X is a random variable denoting red balls. So, X may take value 0, 1, 2, 3 or 4; because we are taking out 4 balls from this box, so that means, this random variable X can take value 0, 1, 2, 3, 4. So, this is red balls; red balls are 6, what will be the P X of this? If we are taking out 0 balls, 0 balls means 0 red balls. So, from 6 red balls we are taking 0 balls, and remaining are how many balls 8; from 8, we are taking out 4; so that means we are taking 4 balls.

Now, from this for 1 red ball; so from 6 red balls we are taking out 1, and from 8 red balls we are from 8 total balls total means blue plus green we are taking out 3 balls. For this 2, it is 6 C 2 and it is 8 C 2; then for X equal to 3, X equal to 3 means, we are taking 3 red balls; 3 red balls out of 6 balls, so this will be; this will be 6 C that means, we are taking out 3 red balls 3 and it is 8 C 1 and it is 6 C 4 8 C 0. And since it is a probability, so it which must be divided by the total cases. Total are 14 balls and out of which we are taking out 4 balls, so it must be divided by 14 C 4, it must be by 14 C 4, 14 C 4, and 14 C 4.

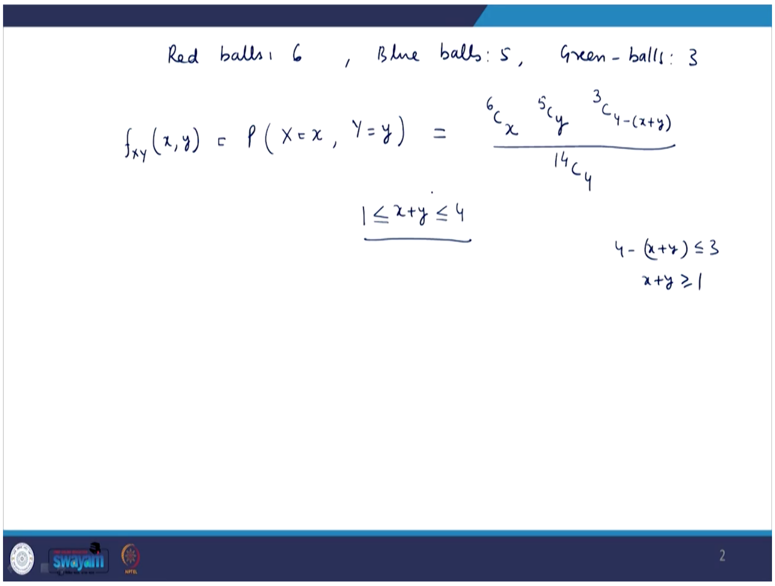
So, this will be the probability mass function for X random variable X, where X is the random variable denoting red balls that means, if X equal to 1 suppose. X equal to 1 means, we are taking 1 red ball; if we are taking 1 red ball out of 6 ball, the total cases will be 6 C 1. And from the 8 remaining balls we are taking out 3 balls, so it is 8 C 3 and divided by the total is 14 C 4. So, this will be the probability for X equal to 1, and similarly other cases we have discussed.

Now, Y denotes the random variable denoting the number of blue balls drawn. So, now it is blue ball; Y it is random variable denoting the blue balls, blue balls. So, again how many blue balls we are having? 5 blue balls we are having, so yes we can draw up to 4; so it is 0, 1, 2, 3, 4. So, what will be P Y?

Now, out of 5 blue balls we are drawing no blue ball 0 that means, 5 C 0 multiplied by remaining balls are 9: 6 plus 3 – 9, so that means, 9 C 4 upon 14 C 4. Here it is 5 C 1 into 9 C 3 upon 14 C 4 by similar concept; 5 C 2 that means, out of 5 blue balls we are drawing 2 blue balls and the remaining are 9 balls out of which we are drawing 2 balls that means, 14 C 4.

Now, it is 5 C 3 9 C 1 14 C 4; it is 5 C 4 9 C 0 14 C 4, so this will be the probability mass function for random variable denoting the blue balls drawn. So, in this way we can we can find out the probability mass function for capital X and capital Y. Now, suppose in the same problem we want to find out the joint probability mass function for X, Y. So, how can we find that? So, let us now we can see.

(Refer Slide Time: 08:46)



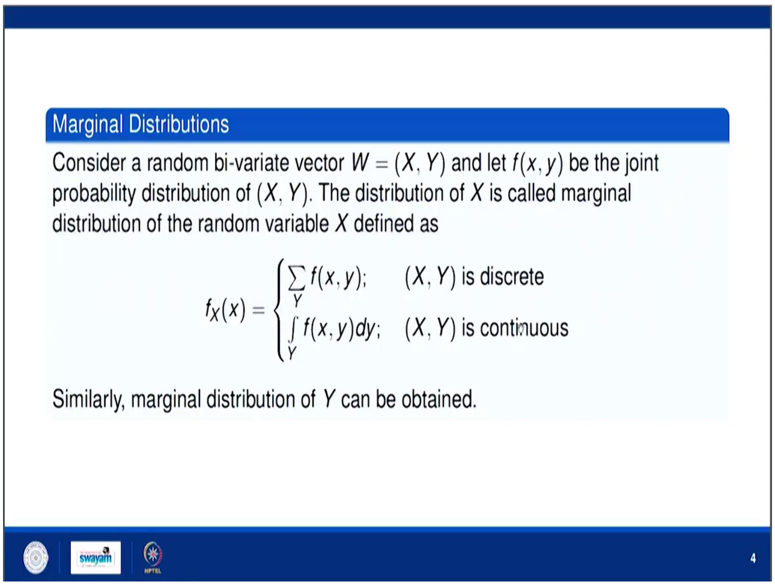
So, red balls are let us write it again, so red balls are how many? Red balls are 6, 5, 3; then blue balls are 5, and green balls are 3. Now, we want to compute the joint pmf of X, Y. So, joint pmf of X, Y; let us denote it by f x, y; so now, it can be calculated as. So, we want to find out we want to take out x, x red balls and y green balls ok.

We want to take out x red balls and y green balls, we can denote it by P also; as we have shown it here, we as we have shown it here P, so we can denote it by P also; let us use the same notations which we have seen here. So, this will be now x x are the number of red balls, so red balls means it is 6 C x ok. So, let us see y the number of blue balls, so blue balls are 5, so it is 5 C y.

So, we have taken out x plus y balls, but we need, but we need 4 balls. So, from the remaining 3 balls, we can find out this much balls, so that so that a total number of balls collected are 4; and total probability will be 14 C. Now, of course X plus y must be must be less than or equal to 4 and must be greater than or equal to must be greater than equal to 2, because see this value can this value the maximum value this can take as 3, so that means 4 minus x plus y should be should be less than equal to 3; so that means x plus y should be greater than equal to 1.

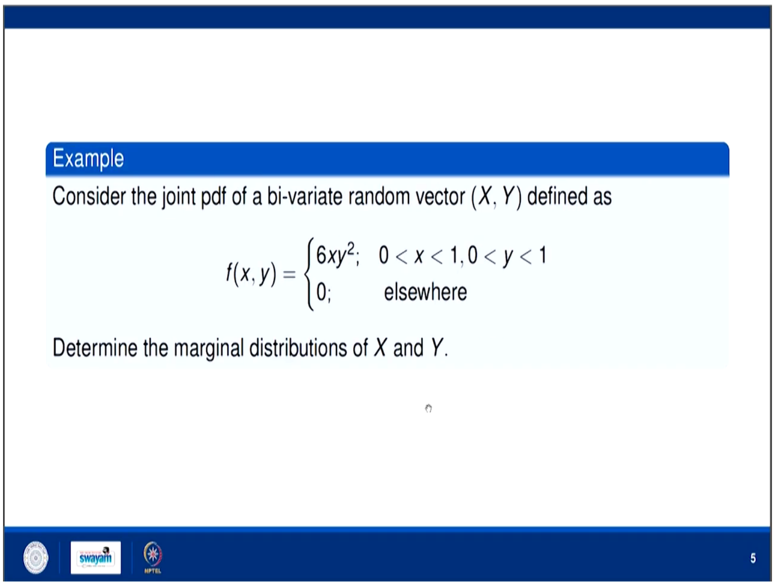
So, it must be greater than equal to one of course, so it must be greater than equal to 1 and it must be less than equal to 4. So, this is the basically joint distribution joint pmf of X and Y. Now, suppose from using this if you want to compute the independent distributions of X or Y or we call it, marginal distributions ok.

(Refer Slide Time: 11:22)



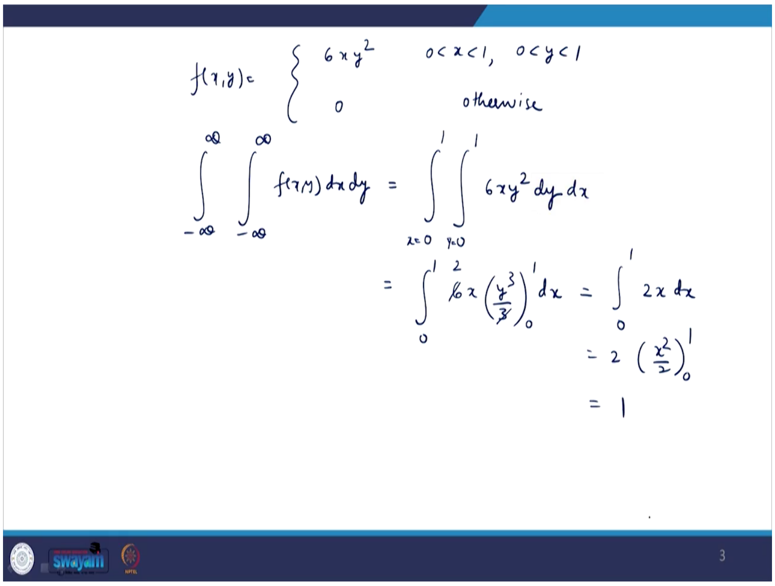
So, let us discuss what marginal distributions are consider a random bi-variate vector X, Y. And let f x, y be the joint probability distribution of X, Y. The distribution of X is called marginal distribution of random variable X defined as, so here we can define basically if X and Y are discrete; then the summing over Y of f x, y, this will give marginal distribution of X random variable X, this we are finding from the joint distribution; this f x, y is nothing but the joint distribution of X, Y. And if X, Y is continuous, so this will be simply integral over Y of f x, y d y; so similarly the marginal distribution of Y can be defined, ok.

(Refer Slide Time: 12:07)



So, let us try to understand the marginal distribution of X and Y, how we can compute using this simple problem. So, consider a joint pdf of a bi-variate random variable X, Y defined as this f x, y equal to this.

(Refer Slide Time: 12:24)

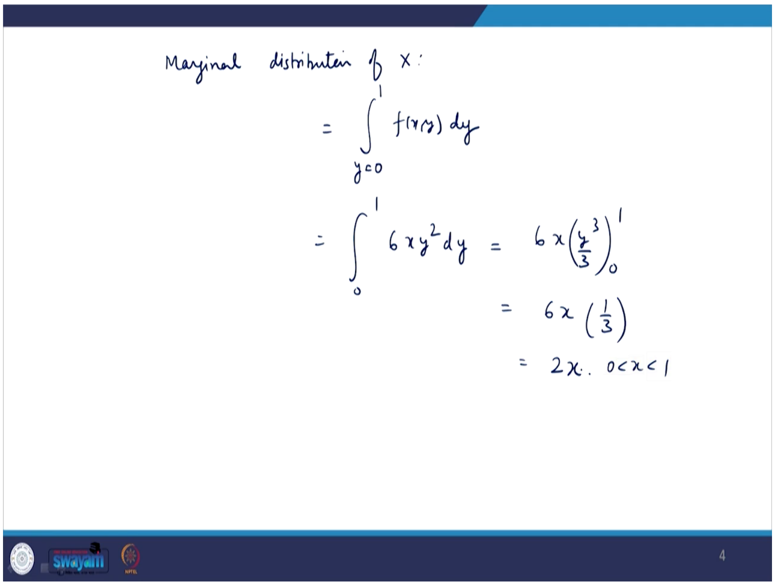


So, here f x, y is what? f x, y is 6 x y square and 0, otherwise. So, of course if you integrate it from minus into plus infinity and minus into plus infinity of f x, y d x d y; it must come out to be 1. So, let us let us check x is varying from 0 to 1, y is varying from 0 to 1; it is 6 x y square d y d x which is equal to integral 0 to 1; so these are limits for y, these are limits for x.

So, it is 6 x y cube upon 3 it is 0 to 1 d x; so this will be equal to this cancels from two times. So, this is integral 0 to 1 2 x d x, which is simply 2 x square upon 2 0 to 1, which is 1. So, yes its come out to be one and of course, it is greater than equal to 0 for every x and y; f x, y is always greater than 0 for every x and y.

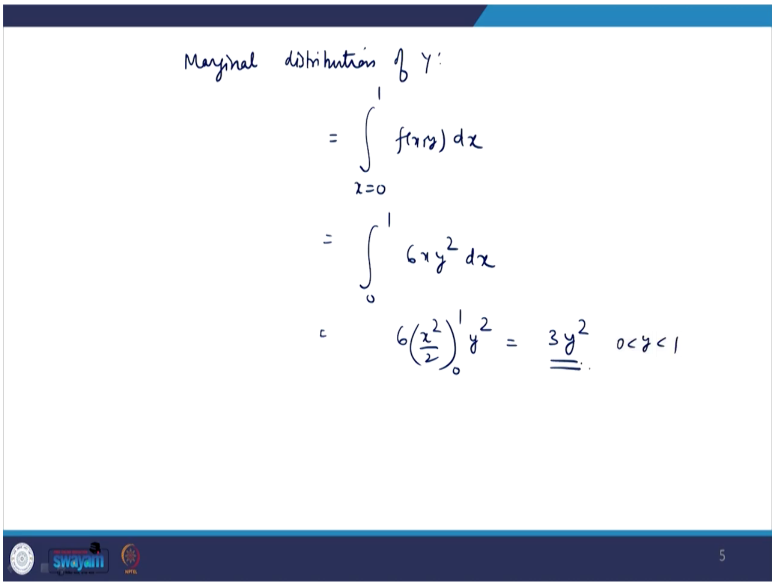
The same properties which is for pmf or pdf or single I mean, for x or y the same will also hold for joint distribution as well that means, the total integral should be 1; and for any x, any y; this f x, y should be greater than equal to 0, so this is joint distribution of X and Y. Now, you want to compute the marginal distribution of X and Y.

(Refer Slide Time: 14:06)



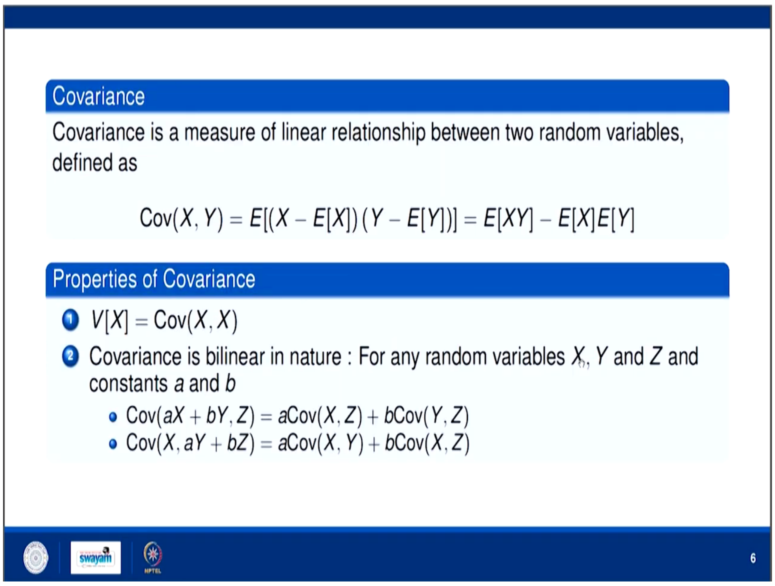
So, let us compute the marginal distribution of X first. So, marginal distribution of X will be so we integrate this f x, y over d y and y is varying from 0 to 1, because it is a continuous function. So, this is equal to this is 0 to 1, this is 6 x y square d x d y; so this is 6 x y cube upon 3 from 0 to 1 and this is 6 x into 1 by 3 which is 2 x. So, this is the marginal distribution of X for this pdf.

(Refer Slide Time: 14:55)



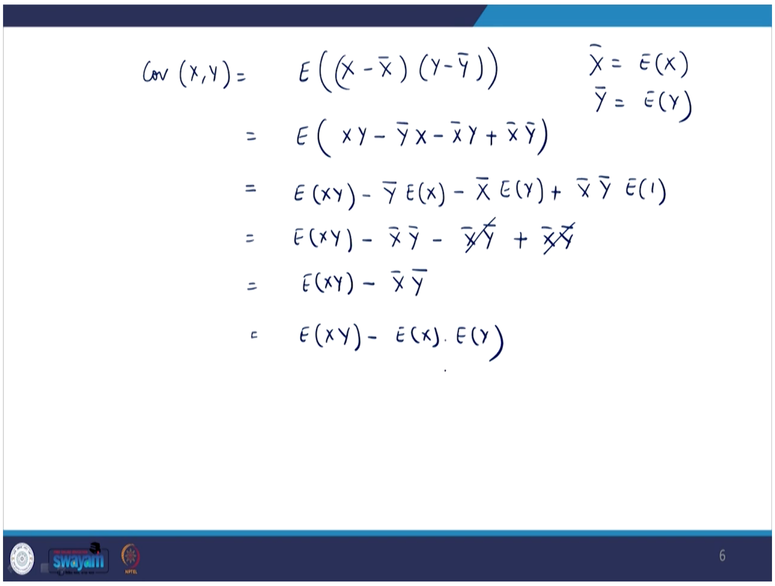
Now, if you want to compute marginal distribution of Y. So, marginal distribution of Y will be integral f x, y not d x x is varying from 0 to 1. So, this is equal to 0 to 1; 6 x y square d x, so this is 6 x square upon 2 0 to 1 y square, so this is nothing but 3 y square. So, this is how we can compute marginal distributions of X and Y; if joint distribution is given to us.

(Refer Slide Time: 15:35)



Now, how can you find covariance. So, what do you mean by covariance first let us discuss, covariance is a measure of linear relationship between two random variables. If you have any two random variables, then it give a linear relationship between them and mathematically it is defined as like this that it is simply expectation of X minus E X into Y minus e Y. Now, now what it comes out to be it is finally comes out to be this expression, let us see how.

(Refer Slide Time: 16:11)



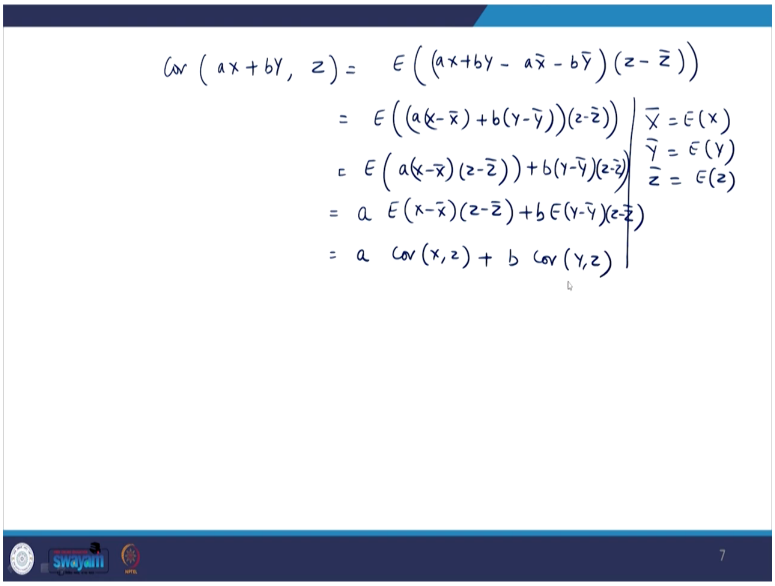
So, what is covariance of X and Y; covariance of X and Y mathematically is expectation of X minus X bar into Y minus Y bar; here X bar is basically E of X and Y bar is basically E of Y. Now, if you simplify this then it is nothing but E of X Y E X Y minus Y bar into X minus X bar into Y minus minus plus X bar Y bar, this is equal to E of X Y.

Now, by the linearity property of expectation we know that this is e of X Y minus now Y bar is of fixed quantity, so it will it can be taken out. So, it is taken out expectation of X minus again X bar is a fixed quantity for this random variable X the mean is fixed. So, it can be taken out and it is expectation of Y plus now, X bar Y bar is totally the fixed quantity and again can be taken out in expectation of 1.

So, this is expectation of X Y minus now E of X is basically expectation of X is X bar, so it is X bar Y bar minus again Y bar X bar plus X bar Y bar. Expectation of 1 is 1 itself, so this is expectation of X Y minus; so one terms cancel out, so it is X bar Y bar. And this is expectation of X Y minus X bar is E X into Y bar is E Y, so in this way we can find out covariance of X or Y.

Now, of course if X, Y equal to X; if they if Y equal to X, then this is nothing but these two values are same; so it is simply whole square and which is nothing but variance of X, so covariance with itself is nothing but variance of X. Next is covariance satisfy these two properties, what are the properties? It covariance of a X plus b Y with Z is nothing but a times covariance of X, Z plus b times covariance of Y, Z. So, it is again very easy to obtain I will try for the first one, the second one can be similarly obtained.

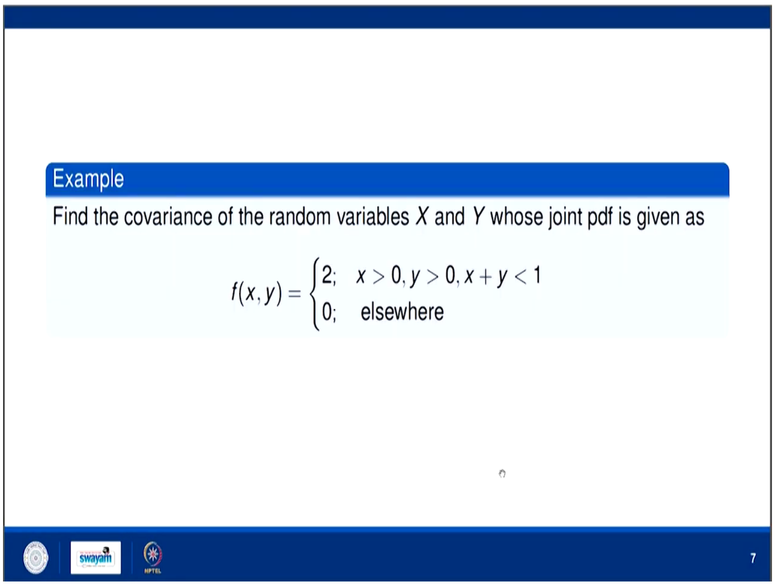
(Refer Slide Time: 18:38)



So, let us try for the first one covariance of a X plus b Y comma Z, so this will be a expectation of a X plus b Y minus a X bar minus b Y bar times Z minus Z bar, as per a definition of covariance of X and Y. Now, here X bar is nothing but expectation of X and Y bar is nothing but expectation of Y and Z bar is nothing but expectation of Z, ok.

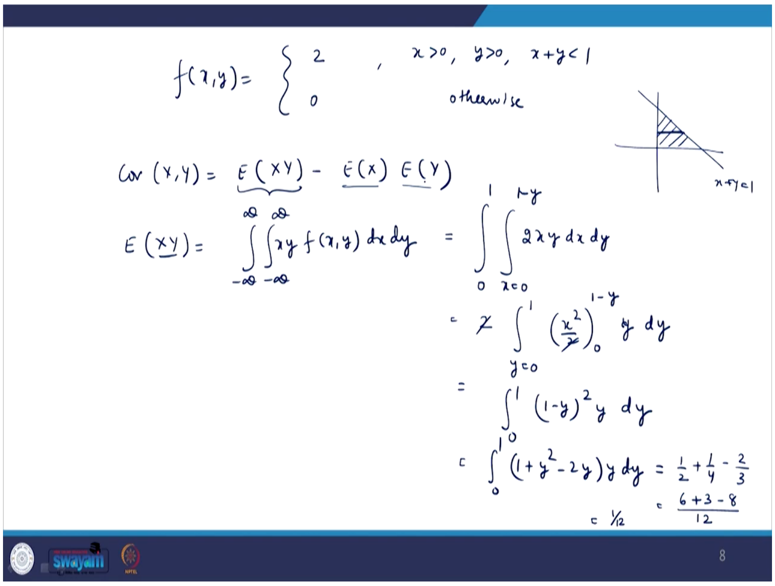
So, it is further equal to expectation of a X minus X bar plus b times Y minus Y bar whole multiplied by Z minus Z bar. So, this is equal to expectations of a X minus X bar into Z minus Z bar ok, this into this plus b times Y minus Y bar into Z minus Z bar. So, this is equal to now by the linearity property of expectation we know that this is equal to a times expectation of X minus X bar into Z minus Z bar plus b times expectation of Y minus Y bar into Z minus Z bar. So, this is a times covariance of X and Z and plus b times covariance of Y and Z. So, other property can also be easily obtained.

(Refer Slide Time: 20:18)



Now, suppose you want to find out the joint the covariance of the random variables X and Y, whose joint pdf is given as this, so how can we find? So, now let us see.

(Refer Slide Time: 20:35)



So, what is what is joint pdf here f x, y is f x y is 2; when x is greater than 0, y is greater than 0 and x plus y is less than 1, and 0, otherwise. So, again one can easily check that it is a pdf by integrating double integral of this in this range, this is equal to 1. And of course, f x, y is greater than equal to 0 for all x and y that is obvious.

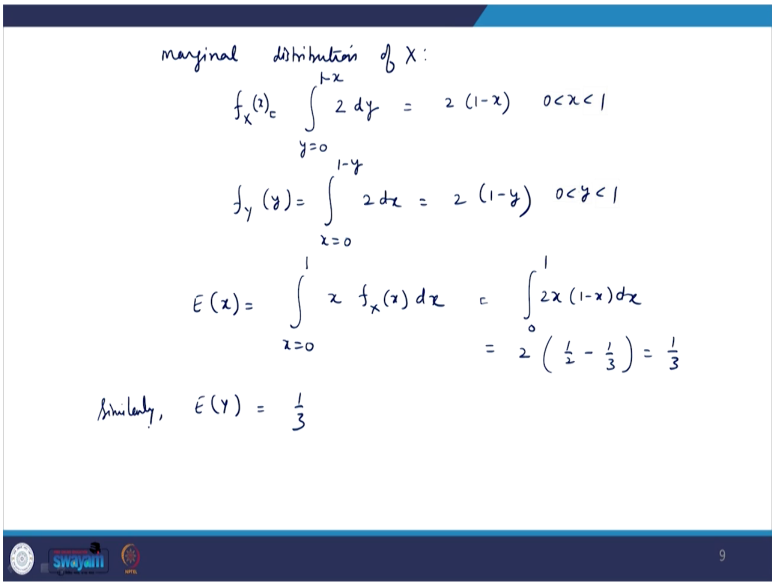
Now, if you want to find out covariance of X and Y for this particular problem. So, this is as per the definition this is expectation of X Y minus expectation of X into expectation of Y. So, first of all we have to compute expectation of X and X into Y. So, expectation of X Y is equal to double integral from I can take it minus infinity to plus infinity minus infinity to plus infinity, then this is f x f x, y x y into f x, y d x d y. I want to compute expectation of X Y, so for that I need joint pdf; so joint pdf is f x y, now this is equal to now it is 2 in this range, otherwise it is 0.

So, what is this range? This is x plus y less than 1, this is x plus y equal to 1, and x plus y x and y greater than equal to 0 means this area. So, if you take it x y and f is 2, suppose we take it d x d y; so we take x d parallel to x axis. So, if you take x d parallel to x axis; what is x here? x here is 0; and what is x here? x here is 1 minus y ok, and y is varying from 0 to 1.

So, this is 2 times integral y from 0 to 1, now you have to integrate x here; x is what? x square upon 2 0 to 1 minus y and it is you have to integrate respect to x; x means x square upon 2 and this is y y d y. So, this is 2, 2 cancels out and this is integral 0 to 1, this is 1 minus y whole square into y d y. This is integral 0 to 1, this is 1 minus 1 plus y square minus 2 y times y d y.

Now, when you integrate it from this you will get y; y means 1 by 2, this is y cube that is 1 by 4 and this is 2 y square that means, minus 2 upon y cube that is 3. So, you can take L C M that is 12 that is 6 plus 3 minus this is 4 into 2 is 8 and that means, 1 upon 12; so this is 1 upon 12, ok. Now, but to compute expectation of X or expectation of Y, first we need marginal distributions of marginal pdf of X and Y.

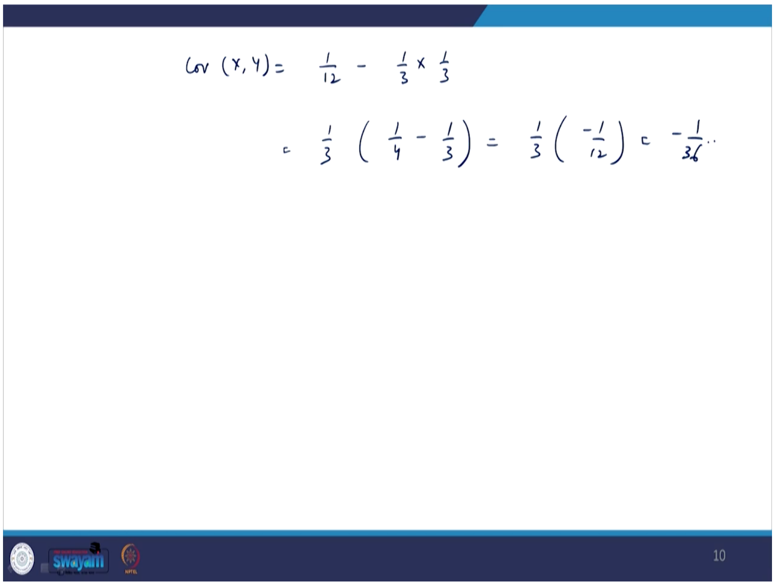
(Refer Slide Time: 23:56)



So, first is marginal distribution of X. So, marginal distribution of X will be it is integral over y and it is integral of y 2, so y is varying from where to where? See in this example, y is varying from 0 to 1 minus x ok, so that means y is varying from 0 to 1 minus x and it is basically d y, ok; so this is equal to 2 1 minus x.

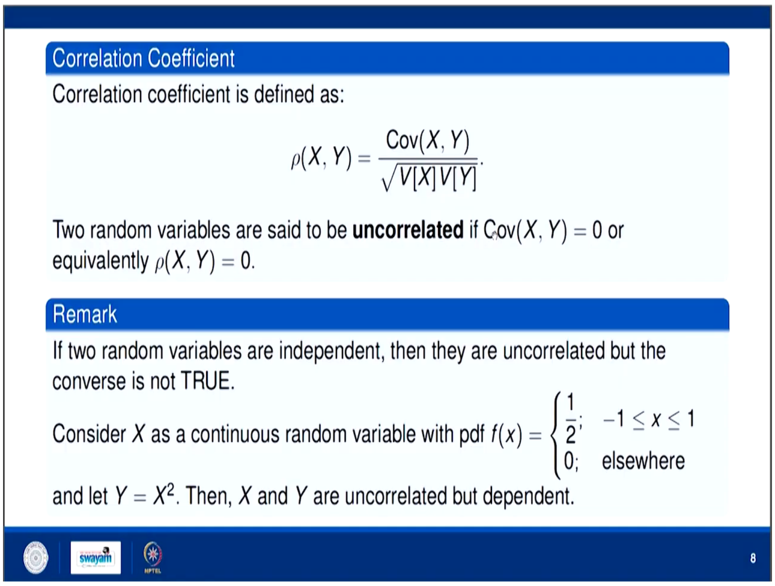
Similarly, marginal distribution of Y will be integral x is varying from 0 to 1 minus y 2 d x, so it is 2 1 minus y. So, expectation of X will be integral integral x into f X x d x and x is varying from 0 to 1 ok. So, this is integral 0 to 1 x into 2 x into 1 minus x d x. So, this will be equal to 2; it is x x means 1 by 2 minus 1 upon 3 that is 1 by 3 ok. And similarly, expectation of Y will be again 1 by 3, because it is symmetrical respect to X and Y, ok. So, what will be covariance of X and Y now?

(Refer Slide Time: 25:38)



So, covariance of X, Y will be equal to expectation of X, Y; 1 upon 12 we have just computed into 1 upon minus 1 upon 3 into 1 upon 3, so that will be the expectation of X and Y; so that we can calculate 1 upon 3 can be taken out common, so it is this value; this is 1 upon 3 into minus 1 upon 12 that is minus 1 upon 36, so that will be the covariance of X and Y. So, in this case they are negatively correlated that means, if X is increasing Y is decreasing or Y is increasing then X is decreasing ok. So, in this way we can find out covariance of two random variables X and Y.

(Refer Slide Time: 26:19)



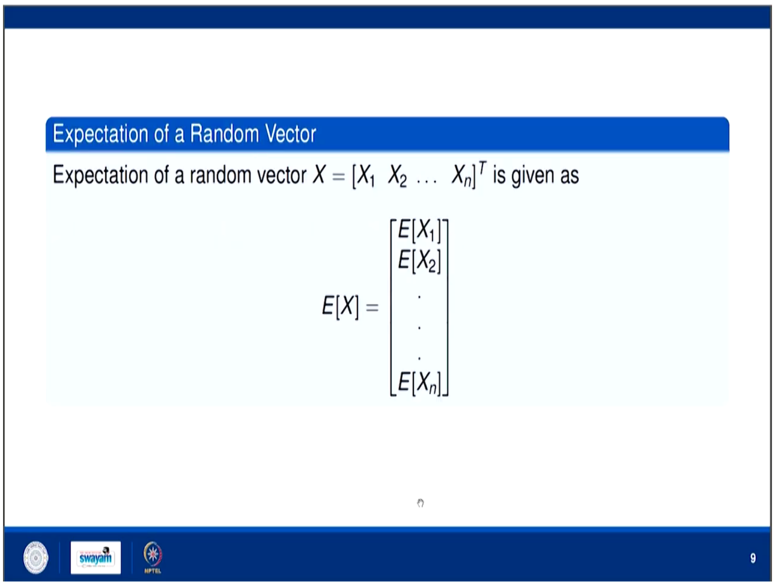
Now, the correlation coefficient, the correlation coefficient of X and Y may be defined as covariance of X, Y divided by standard deviation of X into standard deviation of Y. So, this is a basically a coefficient and if it is 0, two random variables are said to be uncorrelated; if covariance is 0 or covariance correlation coefficient is 0. Now, if two random variables are independent, then covariance between them is 0 or we can say that correlation coefficient is zero, but converts is not true.

See, this is a example which we can easily see. In this example, if you take Y as X square, a random variable Y as taken as X square; so clearly they are not independent they are dependent, but still if you find the correlation coefficient or covariance of X and Y for this particular problem, then it comes out to be 0 or they we can say that they are uncorrelated.

If you find covariance of X and Y which is expectation of X, Y which is by this definition it is the expectation of X, Y minus expectation of X into expectation of Y. Now, Y is X square, so it is expectation of X cube minus expectation of X into expectation of X square. Now, what is expectation of X cube for this problem, for this problem expectation of X cube will be integral minus 1 to plus 1; 1 by 2 x cube d x ok, but x cube is an odd function, so it is directly 0.

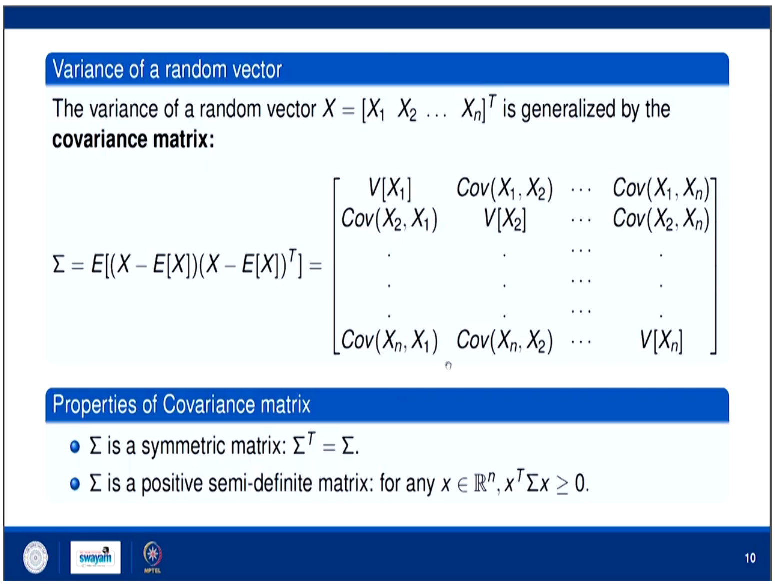
And similarly, expectation of X is again 0, because that is nothing but integral minus 1 to plus 1 X into f x d x which is again an odd function. So, since both the terms since both the terms of this expression this is 0 and this is 0, so we can say that covariance of X and Y is 0. So, they are so they are uncorrelated, but still they are dependent. So, we can say that if they are independent, then correlation coefficient is 0, but converts may not be true.

(Refer Slide Time: 28:39)



Now, if we have a random vector, then we may have expectation of this random vector and that can be defined expectation of X in X 1, expectation of X 2 into up to expectation of X n.

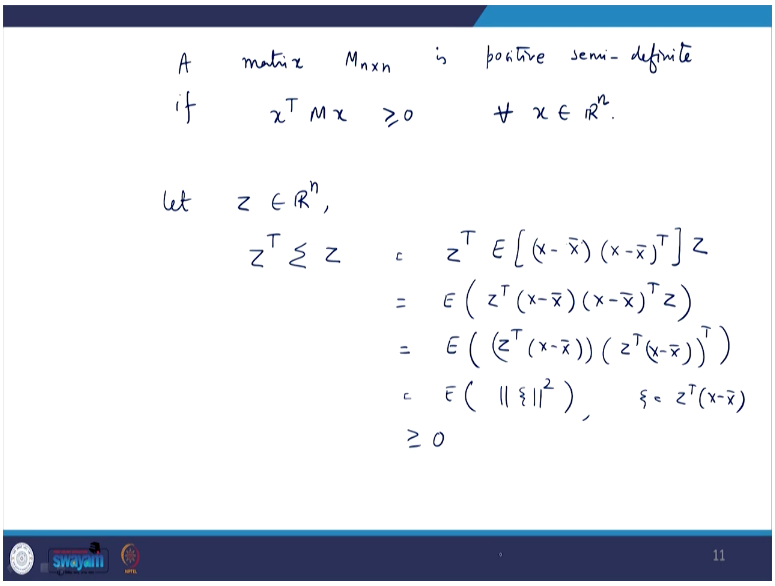
(Refer Slide Time: 28:53)



And in this case, if we are having a random vector, in this case we have covariance matrix. And the covariance matrix can be defined as; this is sigma, this is a notation for covariance matrix and this may be defined as expectation of X minus E X into X minus E X transpose, so that is given as when they are equal. So, this is this gives variance; when it is X 1, it is X 1, it gives variance; when it is X 1, it is X 2, it give covariance of X 1, X 2; when it is X 1, it is x 3, covariance of X 1 X 3 and similarly others.

So, instead of with this basically gives a covariance matrix, ok. And we have a very important property of covariance, this covariance matrix and that is first of all it is symmetric that is clearly seen from the matrix itself, because this matrix equal to its transpose and number 2, this is always positive semi-definite. So, why we are saying it positive semi-definite that is easy to prove, so let us see. So, when a matrix is said to be positive semi-definite?

(Refer Slide Time: 30:08)



So, a matrix M of order n cross n is positive semi-definite; if x transpose M x is greater than equal to 0 for all x belongs to R n. Now, you want to show that this summation matrix or the covariance matrix is positive semi-definite. So, let us take a vector Z, let Z belongs to R n and take Z transpose sigma Z and I want to show that this is greater than equal to 0 for every Z. So, this is Z transpose and this Z is nothing but expectation of X minus expectation of X or we can call it X bar; X bar is nothing but expectation of X into Y minus Y bar transpose into Z, ok.

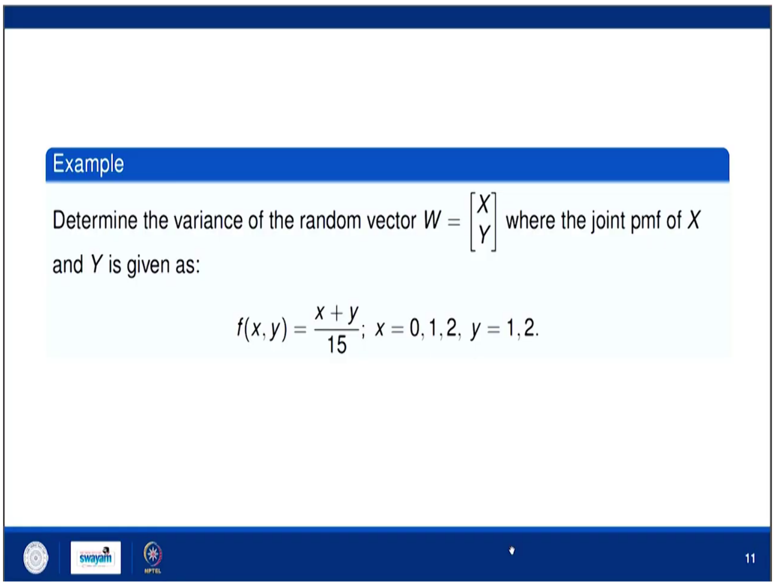
So, so we can see it from here, it is expectation of X in minus E X expectation of E X minus E X, ok; X minus e X again will come, because it is a vector. So, this X is basically X 1, X 2, X

3 up to X n. So, this is further equal to we can because this Z is vector free from X, so we can take it inside; so it is Z transpose X minus X bar into X minus X bar whole transpose into Z, we can we can write it like this. And this is equal to this is equal to expectation of X ok, this is equal to expectation of we can write it like this, Z transpose X minus X bar into Z transpose X minus X bar whole transpose.

So, a vector multiplied by its transpose that is nothing but norm square; this, this is nothing but multiplication of this into this; this is but nothing but Z transpose X minus X bar, ok. Now, we know that if this is positive, then expectation is also positive; so that is greater than equal to 0, because this is always positive, this is non-square is always positive. So, expectation of a positive quantity is always greater than equal to 0, so we have shown that this is greater than equal to 0 that means covariance matrix is always positive semi-definite.

So, in this way suppose two random variables are given to you, and you want to find out the covariance matrix for that; so yes, we can find out. How we can find out? We have to first find the variance of X 1, variance of X 2; covariance of X 1, X 2 which is equal to of course covariance of X 2, X 1; and that 2 cross 2 matrix will be basically covariance of random variable X and X 1 and X 2, ok.

(Refer Slide Time: 33:24)



So, this examples can be solved very easily; if you want to find out covariance of I mean covariance matrix of this two random variables X and Y, whose joint pmf is given as this, so that can easily be obtained that can easily be obtained using this definition.

So, in this lecture we have seen that joint probability distribution is given to you. How can you find out marginal distributions and using that how can you find out covariance of random variables X or Y. Not even two random variables, if you are having random vector containing n random variables that we can also find out the covariance matrix, which is which is having two important properties that is first of all it is symmetric and second it is positive semi-definite.

So, these term we use in machine learning; covariance, how we can find out covariance or how we can find out covariance matrix of a random vector.

So, thank you very much.