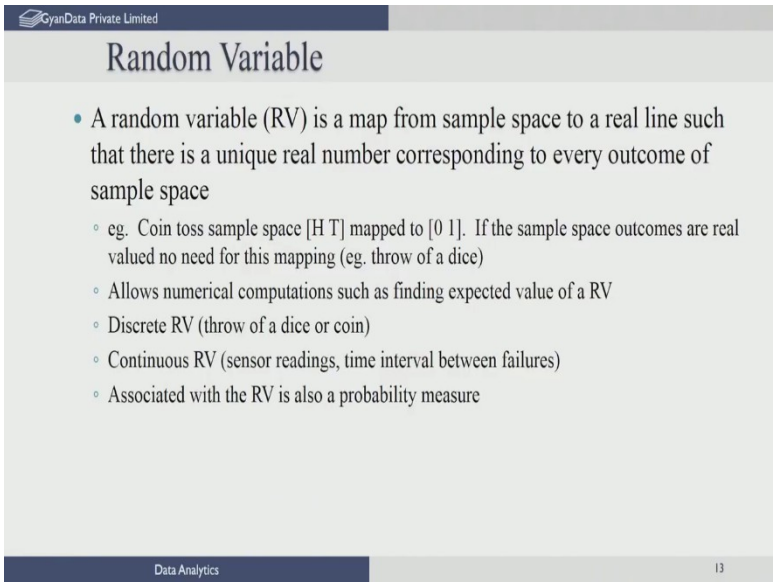
Data Science for Engineers
Prof. Raghunathan Rengaswamy
Department of Computer Science and Engineering
Indian Institute of Technology, Madras

Lecture – 20
Random Variables and Probability Mass/Density Functions

In the previous lecture I introduced the concept of Random Phenomena and how such phenomena can be described using probability measures. In this lecture, I am going to introduce the notion of random variables and the idea of probability mass and density functions. We will also see how to characterize these functions and how to work with them?

(Refer Slide Time: 00:37)



So, a random variable is a function which maps the outcomes of a sample space to a real line. So, there is a unique real number that is associated to every outcome in the sample space. Why do we need a notion of a random variable?

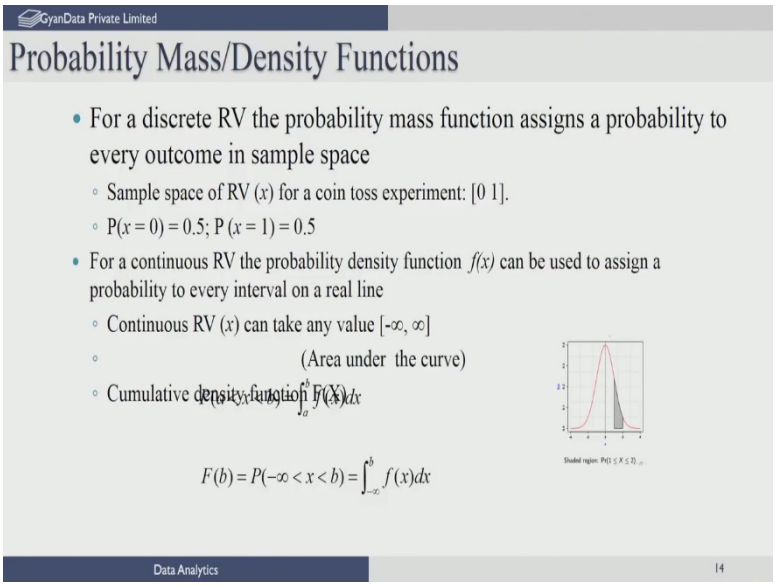
Let us take the example of a coin toss experiment in which the outcomes are denoted by symbols H and T. H refers to the head and T refers to the tail. Unfortunately, we will not be able to do numerical computations with such representation therefore, what we do is to map these outcomes to points on the real line. For example, we map H to 0 and tail to 1. The random variable or function that maps the outcomes

of a sample space to this real life that is what we are referring to as a random variable.

Now, if the outcomes of random phenomena are already numbers such as the true of a dice, then we do not need to do this extra e ort. We can work with the outcomes themselves in that case and call them random variables. We will see how we can do numerical computations once we have such a map and before in a way similar way we have discrete and continuous random phenomena. We will describe discrete random variables that maps functions of discrete phenomena to the real line and continuous random variable which maps outcomes of a continuous random phenomena to intervals in the real life.

So, just as we had a probability measure to characterize outcomes of random phenomena, we will have a similar probability measure that characterizes or is associated with this random variable.

(Refer Slide Time: 02:34)



The notion of a probability mass or a density function is a measure that maps the outcomes of a random variable to values between 0 and 1. For example, if we take the coin toss experiment and associate a random variable x whose outcomes are 0 and 1, then we associate a probability for x = 0.5 and a probability for x = 1 which is another 0.5. Notice that this association should follow the same laws of probability that we described in the last lecture. This is a fair coin and that is why the outcomes are given equal probability.

Now, in the case of a continuous random variable, we define what is known as a probability density function, which can be used to compute the probability for every outcome of the random variable within an

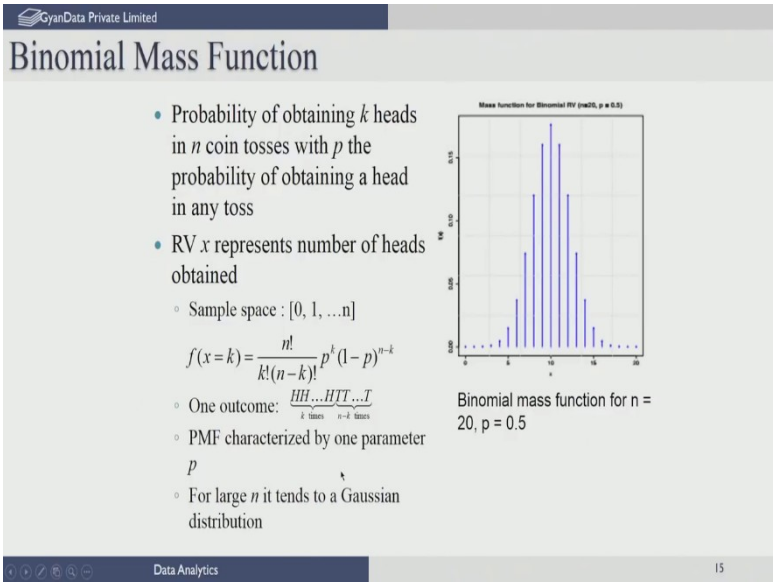
interval. Notice, in the case of a continuous random variable, there are ∞ of outcomes and therefore, we cannot associate a probability with every outcome. However, we can associate a probability that the random variable lies within some finite interval. So, let us call this random variable x which can take any value in the real line from - ∞ to ∞, then we de ne the density function f(x), such that the probability that the variable lies in an interval a to b is defined as the integral of this function from a to b.

So, the integral is an area. So, the area represents the probability. So, this is denoted on the right hand side, you can see a function and here we show how the random variable the probability that the random variable lies between - 1 to town 2 is denoted by the shaded area. That is how we de ne the probability and f(x) which defines this function is called the probability density function.

Again, since this has to obey the laws of probability the integral from - ∞ to ∞ of this function or the area under the entire curve should be = 1 and obviously, the area is non zero therefore it obeys the second law also we actually described in the last lecture. We can also define what is called the cumulative density function, which is denoted by capital F and this is the probability that the random variable x lies in the interval - ∞ to b for every value of b you can compute this value function value and this is nothing, but the integral between - ∞ and b of this density function f(x) dx.

That is known as the cumulative density function for b = - ∞ the cumulative density function value will be 0 and b = ∞ you can verify that the cumulative density function takes the value one. So, this cumulative density function goes from 0 to 1 as the upper limit of the interval goes from - ∞ to + ∞.

(Refer Slide Time: 05:55)



Now, let us look at some sample probability mass functions. Let us take the case of n coin tosses of an experiment where we have do n coin tosses and we are asking the question what is the probability of obtaining exactly k heads in n tosses.

Let us assume p is the probability of obtaining a head in any toss. We also assume that these tosses are independent. So, let us define a random variable x that represents the number of heads that we obtain in these n coin tosses. Now the sample space for x, you can verify goes from takes the value 0 or 1 or n. 0 represents that you observe 0 heads in all the n tosses 1 represents we exactly observe 1 head in n tosses and n represents that all the tosses results in a head.

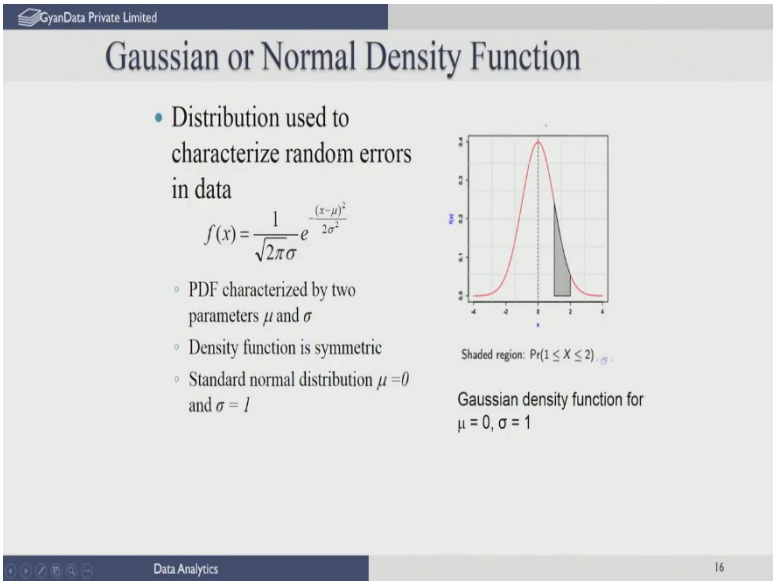
So, clearly each of these outcomes have a certain probability which we can compute and this probability can be computed as follows: let us take the case of k heads in n tosses. Now let us look at one outcome which results in such a such a event. Here I listed the first k tosses as resulting in a head and the remaining n - k tosses resulting in a tail.

Clearly the probability of this particular event is p^k, because I am getting p successive heads and these out tosses are independent. So, p^k is the probability that you will get k successive heads and 1 - $p^{(n-k)}$ would represents the probability that you will get n - k successive tails. So, p^k (1 - $p^{(n-k)}$) represents the probability of this outcome. However, this is only one such outcome of obtaining nk heads. You can rearrange these heads and it is equal into picking k heads out of n tosses. The number of ways in which you can pick k heads out of n tosses is defined by this combination n! divided by k! into(n – k)!.

So, the probability finally, of receiving k heads in n tosses which is denoted by f the random variable taking the value k is given by the probability, which is defined on the right hand side here, this distribution for various values of k. For example, x = 0, x = one can be computed and such a distribution will be called as the probability mass function. For this particular random variable and this particular distribution is called a binomial distribution. As an example of the binomial distribution mass function is shown on the right hand side for n = 20 and taking the probability p = 0.5 clearly, it shows the probability of receiving 0 heads in 20 tosses is extremely small and similarly the probability of obtaining 20 heads in 20 tosses loss is also small as expected the probability of obtaining ten heads has the maximum value as shown.

The vertical line here represents the probability value for a particular number of heads being observed in n tosses. So, clearly since it is a fair coin, we should expect that out of the 20 tosses, 10 heads are most likely to be obtained and this has the highest probability. This distribution is characterized by a parameter p a single parameter p and we can also see for large enough n it tends to this bell shaped curve which is the what is called the Gaussian distribution , which will make use of for large n instead of using the binomial distribution which is computationally more di cult we can approximate by a Gaussian distribution for computational purposes.

(Refer Slide Time: 10:09)



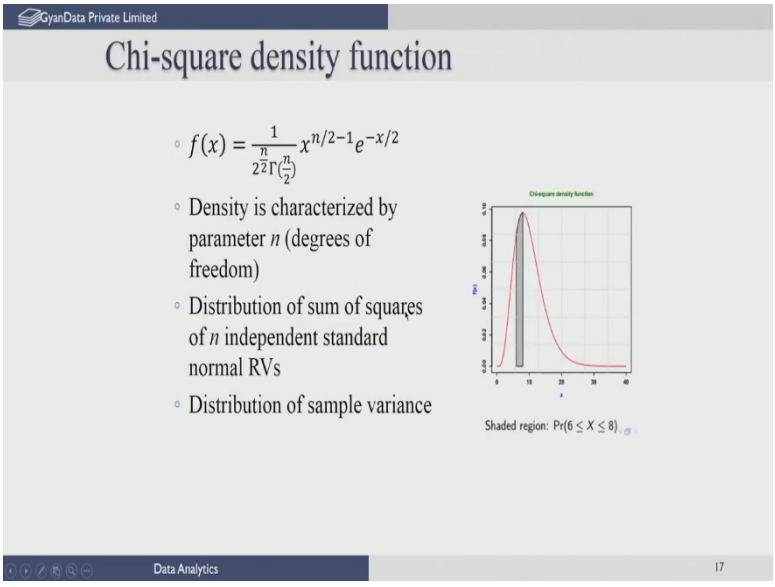
That is an example of a probability mass function of a discrete random variable.

We have now considered a continuous random variable. In this case we will look at what is called the normal density function which is shown on the right. Usually this normal density function is used to characterize what we call random errors in data and it has this density function as given by this $\frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$.

Now, this particular density function has two parameters μ and σ and it has the shape as shown here like a bell shaped curve, which is the normal density function. Notice that it is symmetric and in a particular case of this normal density function is when μ = 0 and σ = 1 and such a normal density function with mean μ =0 and σ =1 is called a standard normal distribution.

Again, if you want to compute the probability of this, that the standard normal variable lies within some interval. Let us say 1 and 2 you have to compute the shaded region. Unfortunately, you cannot analytically do this integration of this function between 1 and 2 you have to use numerical procedures and the our package contains, such functions for computing the probability numerically such that the variable lies within some given interval we will see such computations a little later.

(Refer Slide Time: 11:37)

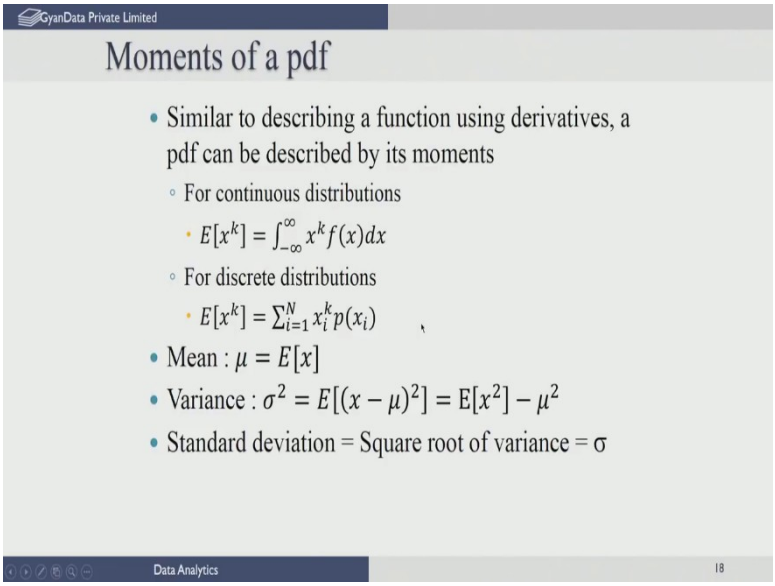


Another continuous random variable who is characterized by density function known as the χ square density function. We do not need to remember the form of this function it has them γ function and so, on. Again this density function is characterized by one parameter which is n called the degrees of freedom. Notice that this function takes values only between the random variable which follows this distribution takes values only between 0 and ∞. The probability to have

negative values is defined to be exactly = 0 and it turns out that this distribution arises when you square a standard normal variable.

So, if you see, the square of a standard normal variable will be a χ square distribution with one degree of freedom. If you take n independent standard normal variables and square and add all of them that will result in a χ square distribution with n degrees of freedom, n representing the number of standard normal variables which you have squared and added to get this new random variable. We can show later that the distribution for sample variance follows a χ square distribution and therefore, it is used to make inferences about sample variances.

(Refer Slide Time: 12:59)



There are other examples of probability density functions such as the uniform density function and exponential density function which we have not touched upon. I would T distribution and. So, on which you can actually look up what we want to do is describe some properties of these density functions.

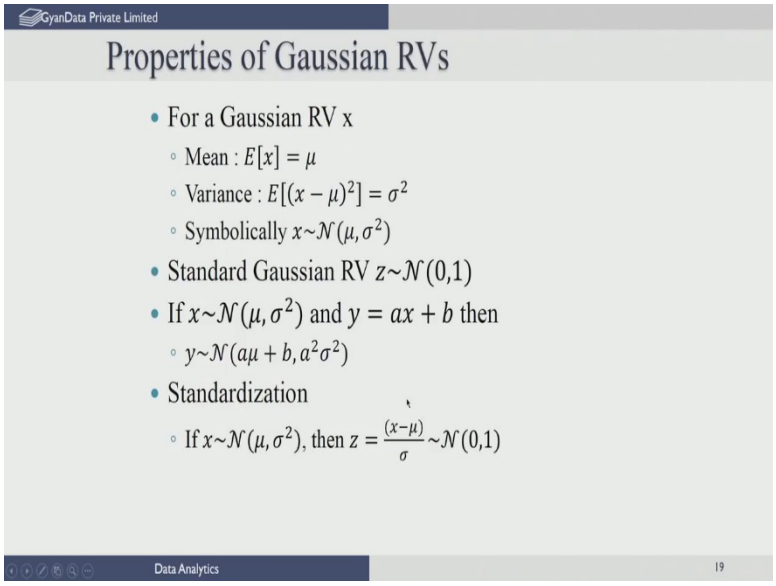
Just as you take a function and talk about properties such as derivatives we can talk about moments of a probability density function. And these moments are described by what is called the symbol expectation e of some function. So, in this particular case, I have taken the function to be x power k and expectation of x power k $E\left[x^k\right] = \int_{-\infty}^{\infty} x^k f(x)\,dx$. This is called the moments of the distribution. If k = 1, you will call it the first moment. If k = 2 you will call the second moment and so on so, forth. So, if you give all the moments of a distribution it is equivalent to stipulating the density function completely.

Typically we will usually specify only 1 or 2 or 3 moments of the distribution and work with them for discrete distributions. You can similarly describe this moment; in this case expectation of x power k is defined as summation -integrations replaced by summation- over all possible outcomes of the random variable x. I represent the outcome and k is the power to which you have raised. So, xi power k probability of obtaining the outcome xi which is very similar to this integration procedure except that the finite number of outcomes and therefore, we have replaced the probability f(x)dx with p (xi) and the value x^k with the outcome xi^k and integral as a summation.

Now there are two important moments that we described for a distribution. What is called the mean or the first moment which is defined as expectation of x and the variance which is defined as denoted by the symbol σ^2 and this is defined as the expectation of x - the first moment which is μ whole square.

This is the function that we want to take the expectation of, which can be obtained by x - μ the whole square f(x) dx. In the case of a continuous distribution we can show that the variance $\sigma^2$is expectation of x squared which is the second moment of the distribution about 0 - μ squared. This proof is left to you, you can actually try to prove this; the standard deviation is defined as the square root of the variance.

(Refer Slide Time: 15:50)



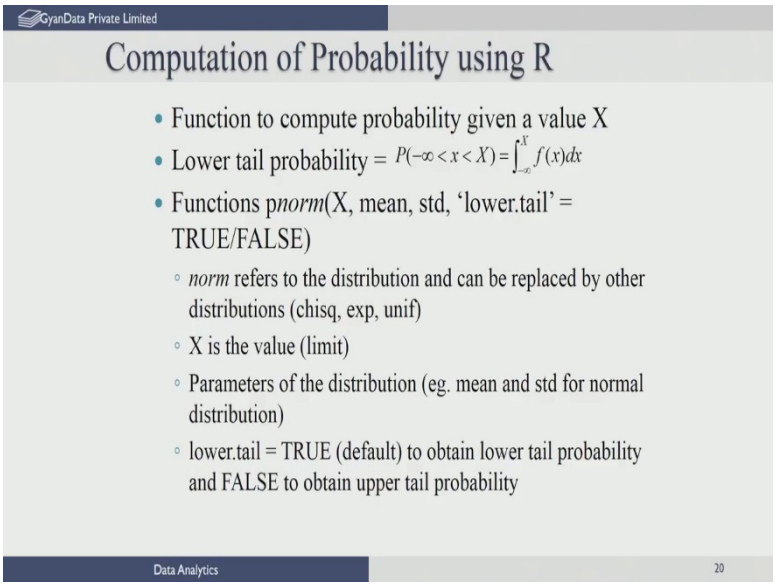
Now, for a Gaussian random variable, if I take the expectation of x which is the mean, we can show that is the first parameter μ in the density function and the variance which is denoted defined as expectation of x - μ the whole squared turns out to be = σ^2which is the second parameter of the distribution.

So, the parameters that we have used to characterize the normal variable is μ the mean and σ^2which is the variance. Typically σ^2 tells you how wide the distribution will be and μ tells me what the value is at which the density function attains the highest probability most probable value. So, μ is also known as the centrality parameter and σ^2 is essentially the width of the distribution tells you how far the values are spread around the central value μ symbolically. We defined this normal distribution variable random variable as distributed as N represents the normal distribution and the parameters are defined by μ and σ^2 which completely defines this density function. A standard normal or standard Gaussian random variable is a particular random variable, which has normally distributed random variable which has mean 0 and standard deviation one and denoted by this symbol.

Now, we can show that if x is a normally distributed random variable with mean μ and σ^2, then if you take a linear function of this x denoted by ax + b, where a and b are some constants, then we can show that y you will also have a normal distribution. But its mean will be our expected value will be a μ + b and its variance would be a squared σ square.

Now we can use this linear transformation to do something called standardization, which you will see often in many many application of hypothesis testing or estimation of parameters, where we simple define if a random variable is normally distributed with mean μ and σ^2 variance. Then we can de ne a new random variable z which is x - μ by σ. That is, you subtract the mean from the variable random variable and divide it by the standard deviation that, this new random variable is a linear transformation or a linear function of x and therefore, we can apply the previous rule to show that z is a standard normal distributed variable which means it has a mean 0 and a standard deviation of variance = 1, this is this process is also known as standardization.

(Refer Slide Time: 18:39)



Now in our there are several functions that allow you to compute probability given a value or the value given the probability. So, we will see some couple of examples of such functions. For example, if you give a value x and ask what is the probability that this continuous random variable lies between the interval - ∞ to this capital x value that you are given then, obviously, I have to perform this integral - ∞ to x of the density function.

And as I said this can only be done numerically and there is a function for doing this it is called pxxx. In the case of a normal distribution you call it pnorm you give it the value upper limit in this case and then you define the mean and standard deviation of the two parameters of the normal distribution and you also specify something; whether you want the upper tail probability or the lower tail probability. We will see what it is and this function will give you the probability value.

This value of this integral, notice this integral is nothing but the area under the curve between - ∞ to x, if lower tail = TRUE it will give you this integral value area between - ∞ and x. On the other hand if lower tail is FALSE, then it will give the in area under the curve between x n ∞. So, x will be taken as the lower limit and ∞ is the higher limit if you say lower dot tail = TRUE it will take excess the upper limit and do the area in under the curve f(x) between - ∞ and x the default value is TRUE. So, this norm can be replaced by other distributions like χ square exponent uniform and so on, so forth to give the probability for other distribution given this value x.

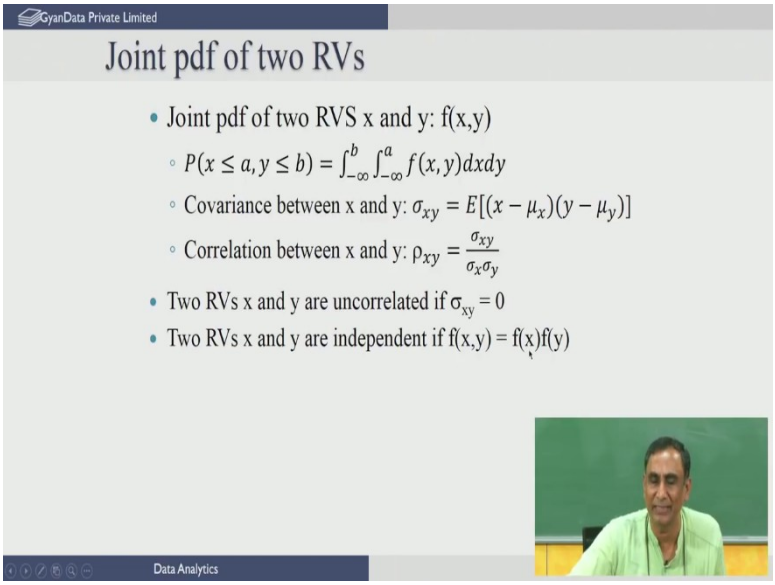
Now, the parameters of the distribution must also be specified for every case in the normal distribution. There is only there are two parameters, but other distributions such as χ squared may have one parameter such as the degrees of freedom and exponent will have one one parameter such as the parameter λ and so on so forth. As I said the lower dot tail tells you whether you want the area to the left of x or to the right of x.

(Refer Slide Time: 21:00)

Then other functions in R one of them is qnorm which actually does what is called the inverse probability; here you give the probability and ask what is the limit X. So, here I have if you give the probability to q norm with the mean and standard deviation parameters of the normal distribution and you say the lower tail = 2, then it will actually compute the value of X such that the integral between - ∞ to X of this density function = the given value p you are specified p and calculating X.

In p norm you are specifying x and computing p. So, this is called the inverse probability function as before if you actually say lower dot trail = FALSE, then this integral will be replaced by x to ∞ such that x to ∞ of f(x) dx = p and it will find the value of x. There are other functions called d norm which computes the density function value at a given x and r norm which are used to generate random numbers from this given distribution.

(Refer Slide Time: 22:04)



Now, having seen the distribution of single random variable, let us move on to the joint distribution of two random variables. In general, we will be dealing with the multivariate distributions which are joint distribution of several random variables, but first we will look at the joint probability density function of two continuous random variables x and y denoted by the function f(x,y).

And the way this density function is used to compute the probability is as before the joint probability that x < = a - ∞ being the other assumed to be the other limit and y < = b or y ranging from - ∞ to b, the joint probability of these two variables lying in these intervals is denoted as computed as the integral - ∞ to b and double integral - ∞ to a f(x) y dx dy.

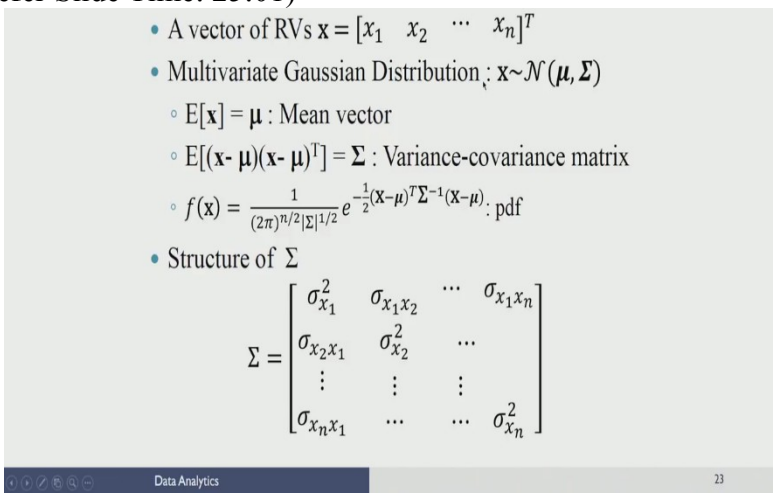
That is the basically the volume of this particular function f(x) comma y. Now when there are two variables you can also de ne other than the variance of x and y which are denoted as $\sigma^2 x$ and $\sigma^2 y$. You can also define what is called the covariance between x and y and this is defined as the expectation of $x - \mu_x$, μ being the mean or expectation of xi multiplied by y - expectation of y which is denoted by μ y this product is the expectation of this product function is defined as the covariance between x and y and denoted by the symbol σ_{xy}.

Now, the correlation between x and y is the standardized or normalized form of this covariance which is nothing but σ_{xy} divided by the standard deviation of x and the standard deviation of y -this is denoted by the symbol ρ_{xy}. We can show that ρ_{xy} varies between - 1 to + 1 depending on the extent of correlation between x and y. Typically, when $\sigma_{xy} = 0$, we $\rho_{xy} = 0$ and we say that the random variables x and y are uncorrelated.

On the other hand, if x and y are independent, then the joint density function of f(x) x comma y can be written as the product of the individual density functions or marginal density functions of x and y. That is f(x) into f of y. This is the extension of this notion of independent variables in terms of probability we defined in the previous lecture where we said the probability joint probability of x and y is basically probability of x into probability of y.

But this is a generalization which defines the notion of independence of two random variables.

(Refer Slide Time: 25:01)



Now, we can now extend this idea of joint distribution of two variables to joint distribution of n variables. Here I have defined the vector x which consists of n random variables x_1 one to xn. And specifically we look at this multivariate normal distribution we denoted by the symbol x multivariate normal with mean vector μ and covariance matrix σ. Now each of these xi components x_1, x_2 and so on have their respective means. If you put them in the vector form, we get this mean vector symbolically written as expectation of x which is a multi dimensional integral, we get this value μ which is known as the mean vector. And similar to the variance we defined what is called the covariance matrix which is defined as expectation of x about the mean μ or x - μ into $(x - \mu)^T$. Remember this is a matrix of variables because x is a vector and if you take the expectation of each of these elements of this matrix you will get this matrix called the variance covariance matrix.

In fact, we can write the multi dimensional normal distribution also has a very similar form if you look at it 1 by 2 π we had square root of 2 π. In this case it will be 2 π n by 2 n represents number of dimension of this vector and we had σ in this case. We have the square root of this matrix covariance matrix σ and we had exponents - half we had a quadratic form. In this case the quadratic form is $(x - \mu)^T \sum^{-1} (x - \mu)$ which is similar to $(x - \mu)^2$ divided by σ^2.

So, it has very similar form we do not need to know the form. We need to know how to interpret μ and σ. And if you look at the structure of σ you will find that it is a square matrix with the diagonally elements representing the variance of each of the elements that is $\sigma^2 x_1$ is the variance of x_1 and $\sigma^2 x_2$, the variance of x_1 and so on, so forth. And the of diagonal elements representing the covariance between x 1 and x_2 or x_1 and x_3, x_2 and x_3 and so on ,so forth pair wise covariance.

Those are the o diagonal elements for example, $\sigma x_1, x_2$ represents the covariance between x_1 and x_2 $\sigma x_1, x_n$ represents the covariance between x_1 and x_n this particular matrix is symmetric and we completely characterized the multivariate normal distribution by specifying this mean vector μ and the covariance matrix σ.