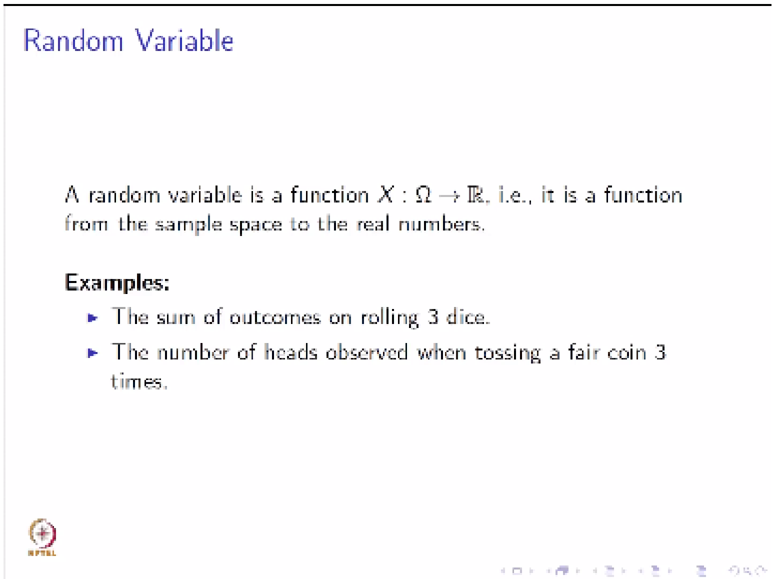
NPTEL

NPTEL ONLINE CERTIFICATION COURSE

(Refer Slide Time: 00:16)

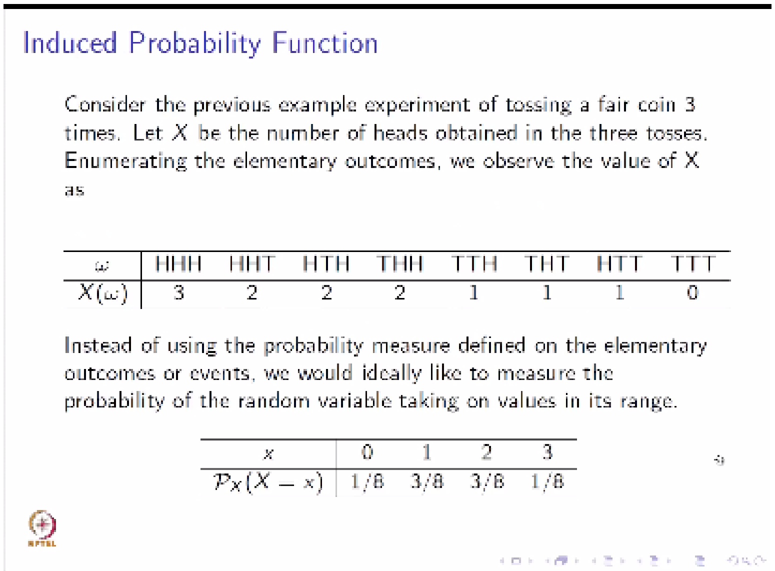


One of the important concepts in probability theory is that of the random variable. A random variable is a variable whose value is subject to variations that is a random variable can take on a set of possible different values each with an associated probability. Mathematically a random variable is a function from the sample space to the real numbers. Let us consider some examples suppose we conduct an experiment in which we roll three dice and are interested in the sum of the outcomes.

For example the sum of files can be observed if two of the dice show up to each and the other die shows up as one. Alternatively the sum of five can also be observed if one die shows up as three and the other two dice show up one each. Since we are interested in only the sum and not the individual results of the dice rolls we can define a random variable which maps the elementary outcomes that is the outcomes of each die roll to the sum of the three rolls.

Similarly in the next example we can define a random variable which counts the number of heads observed when tossing a fair coin three times. Note that in this example the random variable can take values between 0 and 3 whereas in the previous example the range of the random variable is between 3 and 18 corresponding to all dice showing up one and all dice showing up six.

(Refer Slide Time: 01:44)

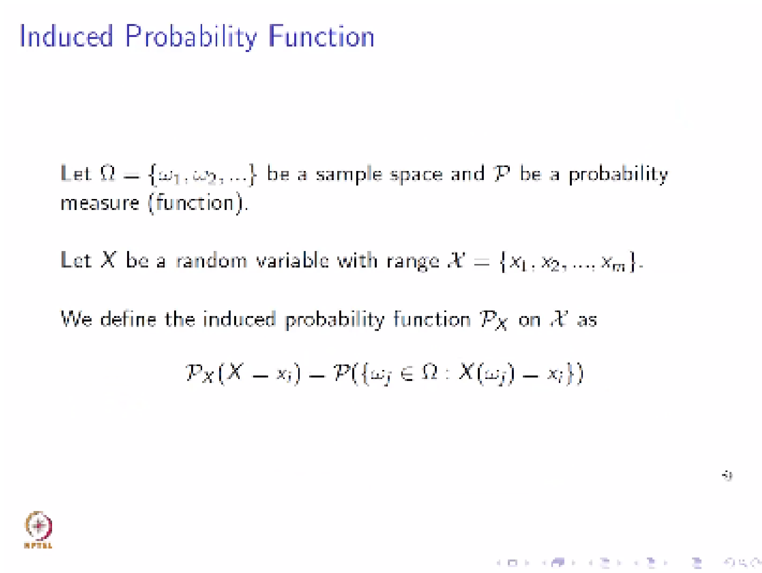


Consider the previous example experiment of tossing a fair coin three times. Let X be the number of heads obtained in the three tosses that is X is a random variable which maps each elementary outcome to a real number representing the number of heads observed in that outcome this is shown in the first table. The first row lists out each elementary outcome and the second row this out the corresponding real number value to which that elementary outcome is mapped that is the number of heads observed in that outcome.

Now instead of using the probability measure defined on the elementary outcomes or events we would ideally like to measure the probability of the random variable taking on values in its range. What we are trying to say here is that when we define probability measure we were associating each event that is subset of the sample space with a probability measure. When we consider random variables the events correspond to different subsets of the sample space which mapped to different values of the random variable.

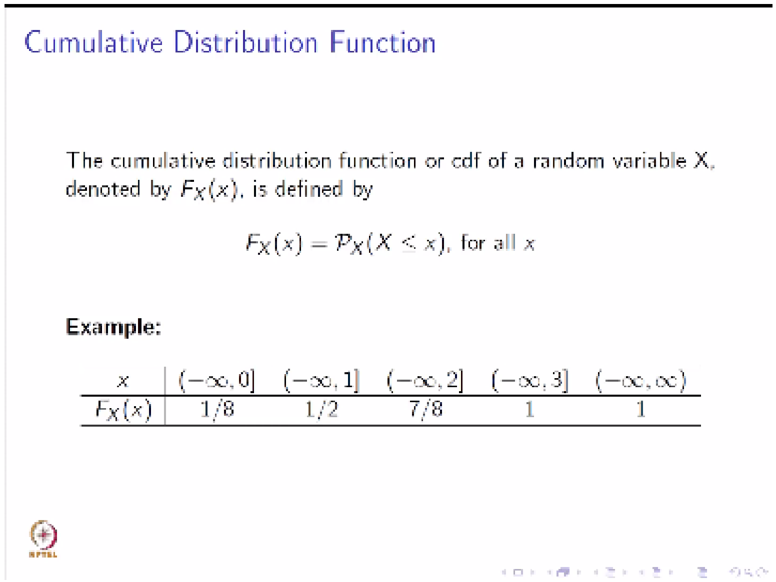
This is illustrated in the second table, the first row lists out the different values that the random variable X can take and the second row lists out the corresponding probability values assuming that the coin toss is a fair coin. This table describes the notion of the induced probability function which maps each possible value of the random variable to its associated probability value. For example, in the table the probability of the random variable taking on the value of 1 is given as 3/8. Since there are three elementary outcomes in which only one head is observed and each of these elementary outcomes has a probability of 1/8.

(Refer Slide Time: 03:30)



From the previous example we can define the concept of the induced probability function. Let ω be a sample space and P be a probability measure, let X be a random variable which takes values in the range X1 to XM the induced probability function px on X is defined as PX, x = xi equals to the probability of the event comprising of the elementary outcomes ωj such that the random variable X map ωj to the value Xi.

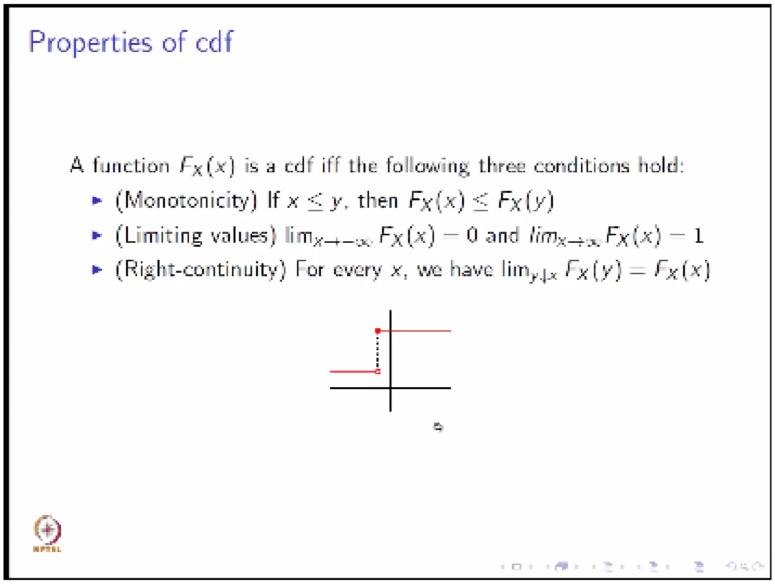
(Refer Slide Time: 04:10)



The cumulative distribution function or CDF of a random variable X denoted by Fx(x) is defined by Fx(x) equals to the probability of the random variable taking on a value less than or equal to x for all values of x. For example, going back to the previous random variable which counts the number of heads observed in three tosses of a fair coin. The following table shows the intervals corresponding to the different values of the random variable X along with the corresponding values of the cumulative distribution function.

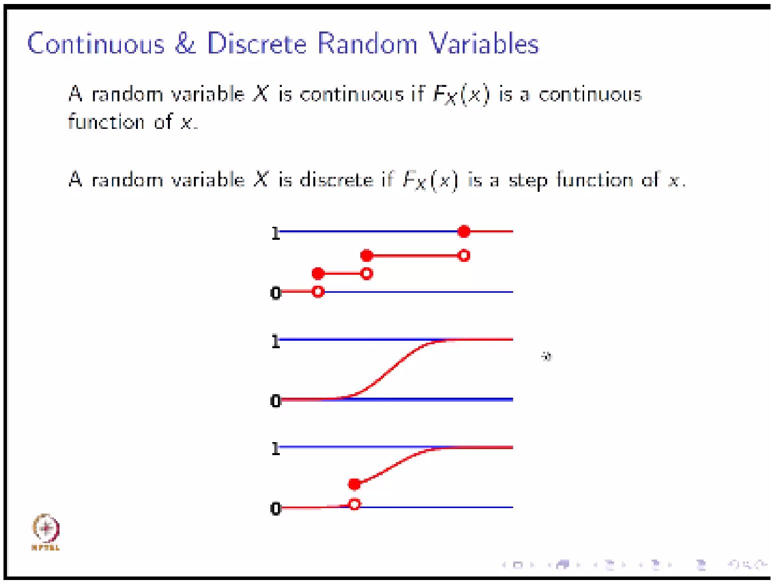
For example, Fx(x)=Fx(1)=1/2, because the probability that the random variable X has a value of one let us just go back to the previous example right the probability that the random variable X has a value of 1 is 3/8, the probability of X that the random variable x=1/8. And therefore, the probability that the random variable X takes on a value with less than or equal to 1 is 1/8+3/8=4/8 or ½.

(Refer Slide Time: 05:28)



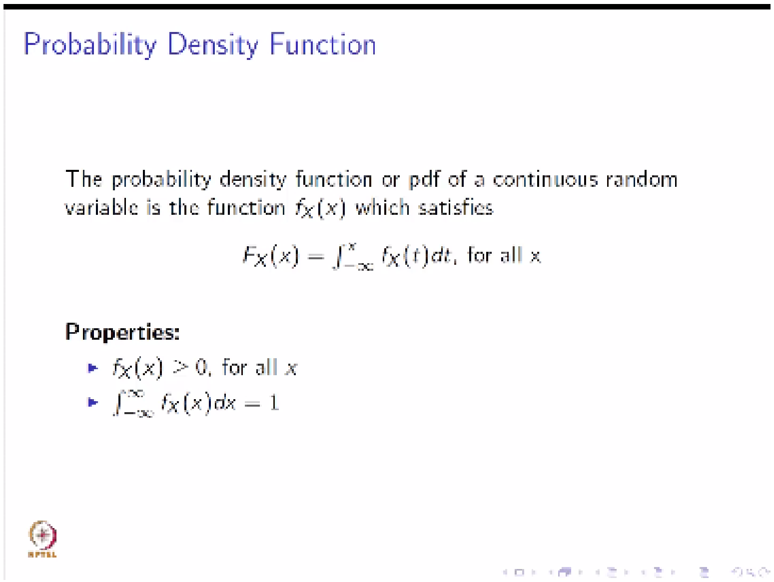
A function is a valid cumulative distribution function only if it satisfies the following properties. The first property simply states that the cumulative distribution function is a non decreasing function. The second property specifies the limiting values, limit X tends to -∞ Fx(x) =0 and limit X tends to ∞ Fx(x) =1. The third property specifies right continuity that is no jump occurs when the limit point is approached from the right this is also shown in the figure below.

(Refer Slide Time: 06:07)



A random variable X is continuous if its corresponding cumulative distribution function is a continuous function of X, this is shown in the second part of the diagram. A random variable X is discrete if its CDF is a step function of X this is shown in the first part of the diagram. The third part of the diagram shows the cumulative distribution function for a random variable which has both continuous and discrete parts.

(Refer Slide Time: 06:33)

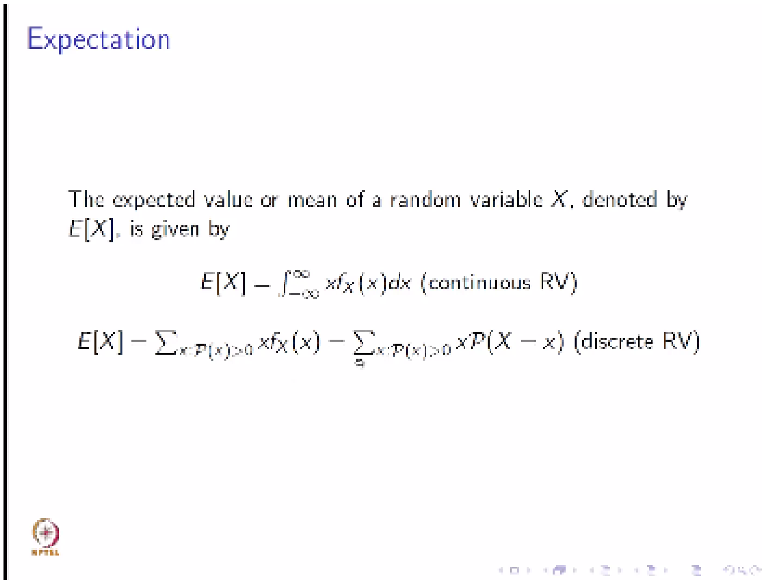


The probability mass function or pmf of a discrete random variable X is given by Fx(x) equals to probability of X = x for all values of x. Thus for a discrete random variable the probability mass function of that variable gives the probability that the random variable is equal to some value. For example, for a geometric random variable X with parameter P the pmf is given as Fx(x) = $(1-p)^{x-1}p$ for the values of x = 1, 2 and so on. And for other values of x the pmf = 0.

A function is a valid probability mass function if it satisfies the following two properties. First of all the function must be non-negative, secondly the summation overall X the value of the function summed over all values of x should be equals to 1. For continuous random variables we consider the probability density function. The probability density function or pdf over and continuous random variable is the function Fx(x) which satisfies the following.

The integral from -∞ to X Fx(t)dt is equals to the cumulative distribution function at the point X. Similar to the pmf the probability density function should also satisfy the following properties. First of all the probability density function should be non-negative for all values of x. Second integrating over the entire range the probability density function should sum to 1.

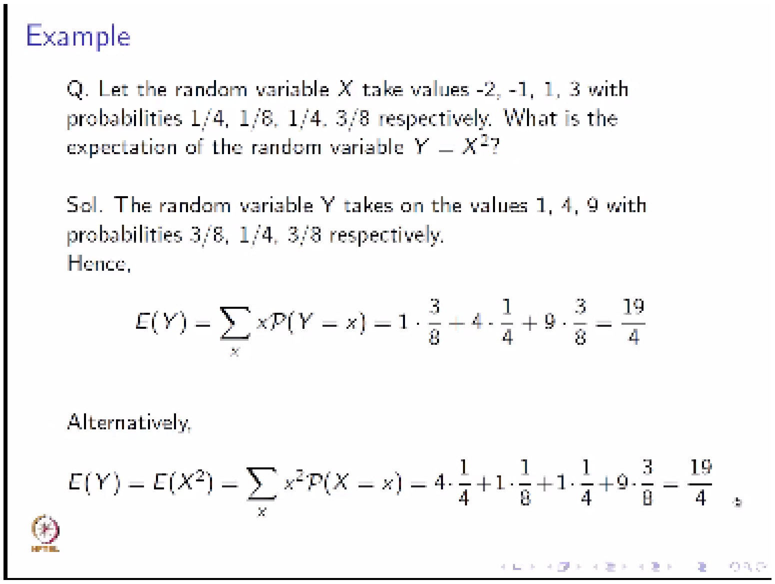
(Refer Slide Time: 08:21)



Let us now look at expectations of random variables the expected value or mean of a random variable X denoted by expectation of X is given by integral -∞ to ∞ x into Fx(x)dx. Note that Fx(x) here is the probability density function associated with random variable x. This definition holds when x is a continuous random variable. In case that x is a discrete random variable we use the following definition expectation of x is equal to sum over all x such that probability of x greater than 0.

That is we consider all values of the random variable for which the associated probability is greater than zero x into Fx(x). Here Fx(x) is the probability mass function of the random variable X which essentially gives the associated probability for a particular value of the random variable thus leading to this definition.

(Refer Slide Time: 09:17)

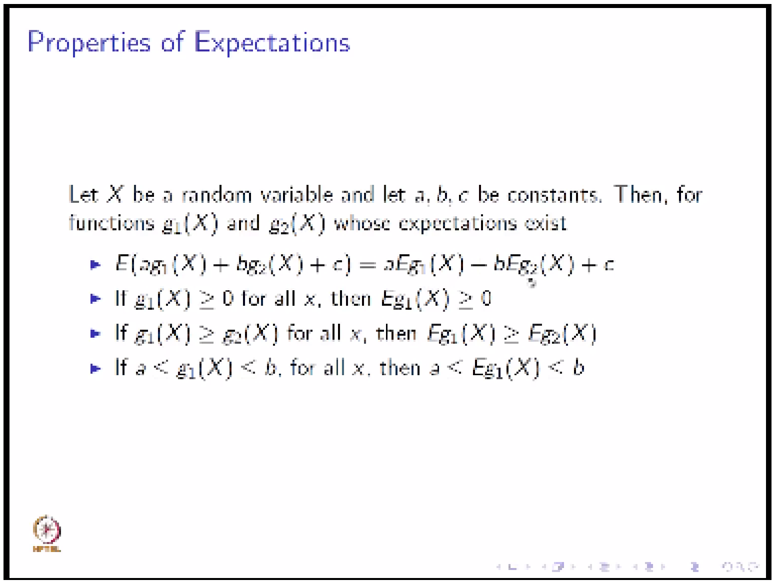


Let us now look at an example in which we calculate expectations that the random variable X take values -2, -1, 1 and 3 with probabilities ¼, 1/8, ¼ and 3/8 respectively. What is the expectation of the random variable $Y\text{-}X^2$? So in this question we are given one random variable the values which this random variable takes and its associated probabilities, but we are interested is in the expectation there are a random variable Y which is defined as $Y\text{-}X^2$.

So what we can do is we can calculate the values that the random variable Y takes along with associated probabilities, since we are aware of the relation between Y and X. Thus we have Y taking on the values 1, 4 and 9 with probabilities 3/8, 1/4, and 3/8 respectively given this information we can simply apply the formula for expectation and calculate the expectation on the random variable Y this is as follows giving a result of 19 /4 another way to approach this problem is to directly use the relation $Y= x^2$ in calculating expectation does expectation of y is simply the expectation of the random variable x^2

So in place in the formula for expectation instead of substituting X we substitute X^2 thus we have some overall x x^2 into probability of x = x calculating the values we get the same answer of 19 / 4.

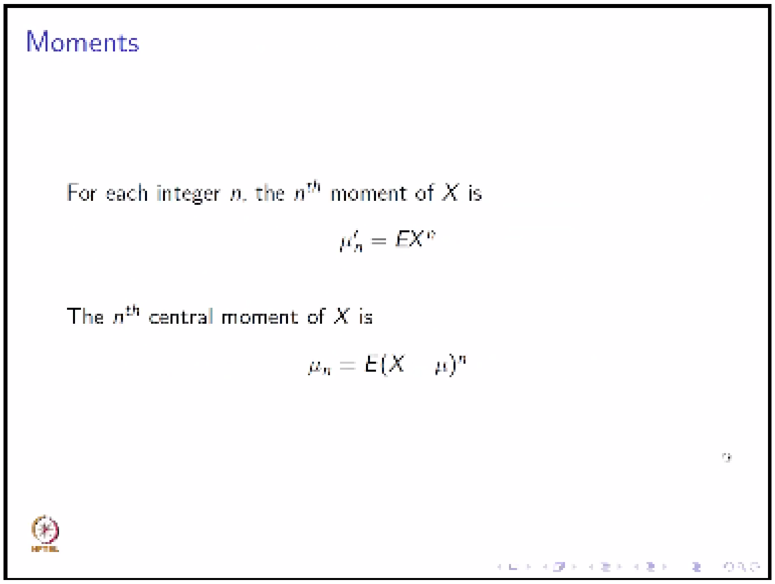
(Refer Slide Time: 10:54)



Let us now look at the properties of expectations Let X be a random variable A B and C are constants and g1 and g2 are functions of the random variable X such that their expectations exist that is they have finite expectations according to the first property expectation of a into g1(x) + b times e2(x) + c = A times expectation of g1 of x + b times expectation of g2(x) +c this is called the linearity of expectations there are actually a few things to note here first of all expectation of a constant is equal to the constant itself expectation of a constant times the random variable is equal to the constant into the expectation of the random variable and the expectation of the sum of two random variables can also be represented as the sum of the expectations of the do random variables.

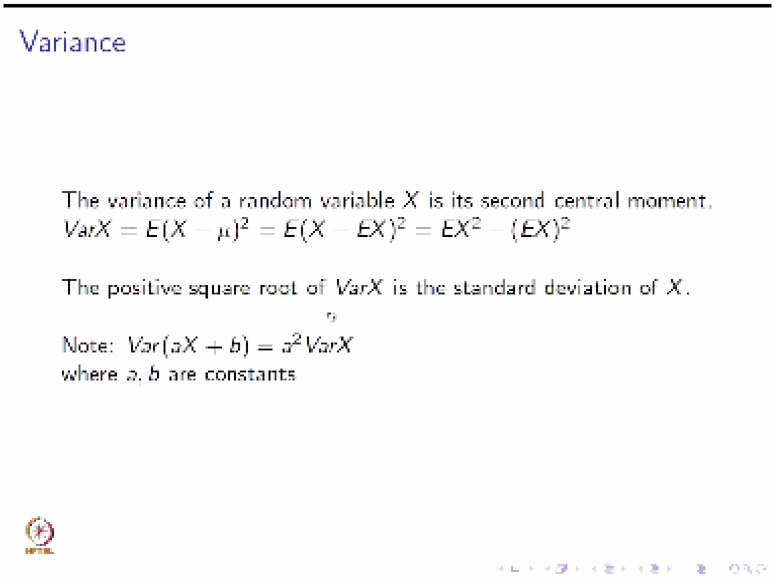
Note that here the two random variables need not be statistically independent according to the next property if a random variable is ≥ 0 at all points then the expectation is also expectation of that random variable is also ≥ 0 similarly if one random variable is > another random variable at all points then the expectation of those random variables also follow the same constraint finally if a random variable has values which are which lie between two constants then the expectation of that random variable will also lie between those two constants.

(Refer Slide Time: 12:34)



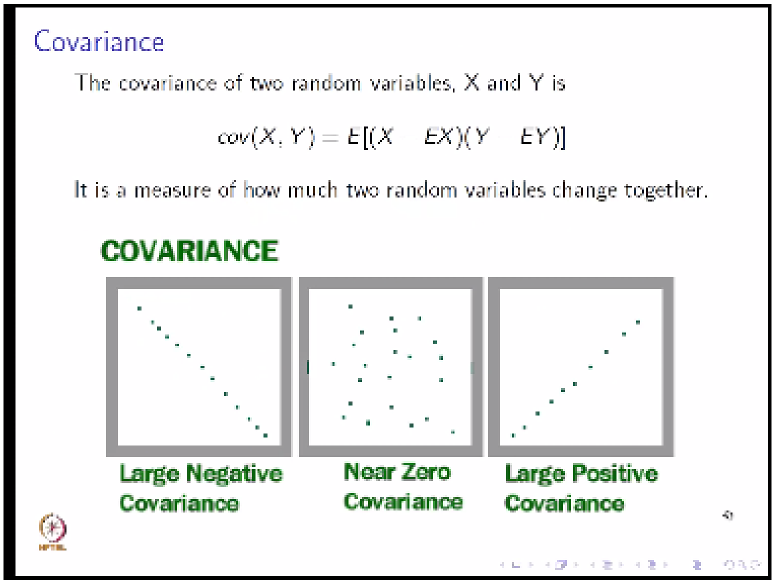
Let us now define moments for each integer n the n^{th} moment of x is μn` = expectation of X raise to the power n also the $n^{th}$ central moment of X is $\mu n$` = expectation of X – μ^n so the difference between moment and central moment is in central moment we subtract the random variable by the mean of the random variable or expected value the two moments that find most common use are the first moment which is nothing but μ` = expectation of X that is the mean of the random variable X and the second central moment which is μ_2 = expectation of X –μ^2 which is the variance of the random variable X.

(Refer Slide Time: 13:23)



Thus the variance of a random variable x is a second central moment variance of x equals to expectation of $X - \mu^2$ note that μ is just the first movement which can be so it can be replaced by expectation of X thus we have variance of X = to expectation of X - expectation of X^2 by expanding disturb and applying linearity of expectations we will finally get variance of x = to expectation of x^2 - x^2 of the expectation of X the positive square root of variance of X is the standard deviation of X note that the when calculating variance the constants act differently and compared to the linearity of expectation. This is a very useful relation to remember variance of ax +b = a^2 into variance of X where A and B are constants.

(Refer Slide Time: 14:17)

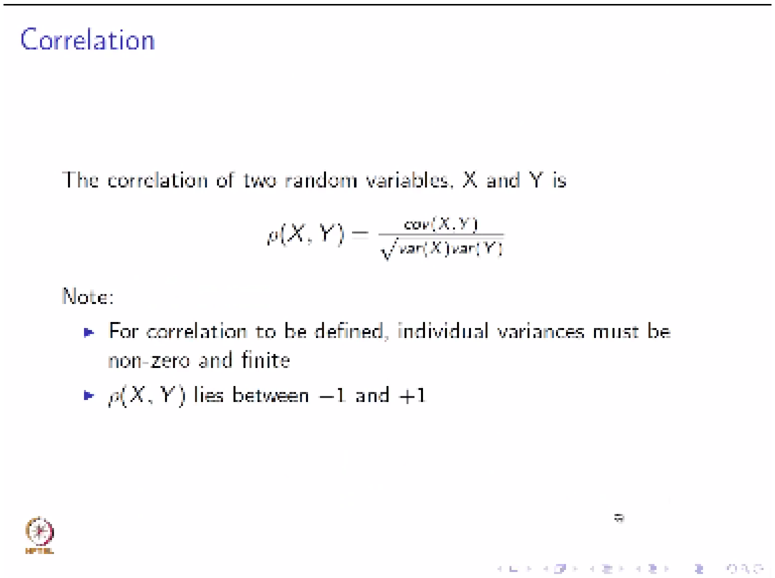


The covariance of two random variables X and Y is covariance of x , y equals to expectation of X - expectation of X x Y - expectation of Y remember that the variance of our random variable X is nothing but the second central moment thus the variance of a random variable measures the amount of separation in the values of the random variable when compared to the mean of the random variable for covariance the calculation is done on a pair of random variables and it measures how much two random variables change together consider the diagram below in the first part assume that the random variable X is on the x-axis and the random variable Y is on the y-axis.

We note that as the value of x increases the value of y seems to be decreasing thus for in for this relationship we will observe a large negative covariance similarly in the third part of the diagram we can see that as the variable value of variable x increases, so does the value of the variable y thus we see a large positive covariance however in the middle diagram we cannot make any such statement because as x increases there is no clear relationship as to how Y changes thus this kind of a relationship will give zero covariance.

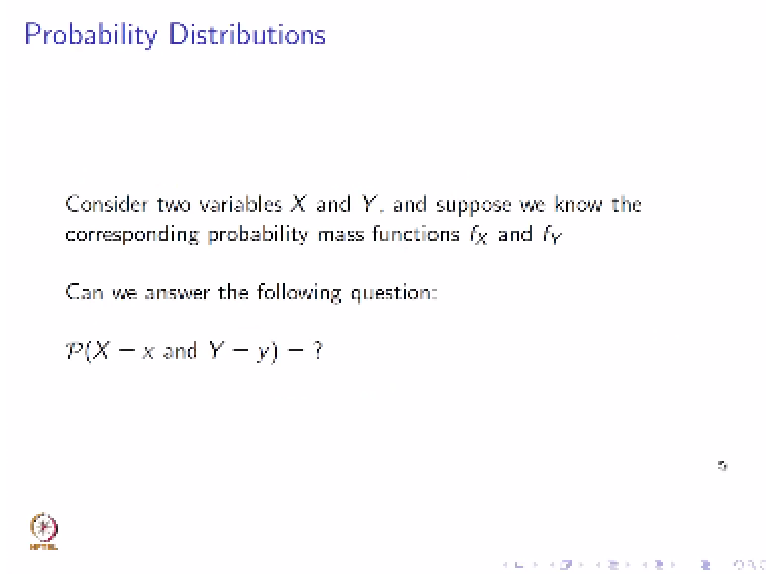
Now from the diagram it should immediately be clear that covariance is a very important term in machine learning because we are often interested in predicting the value of one variable by looking at the value of another variable we will come to that in further classes.

(Refer Slide Time: 16:04)



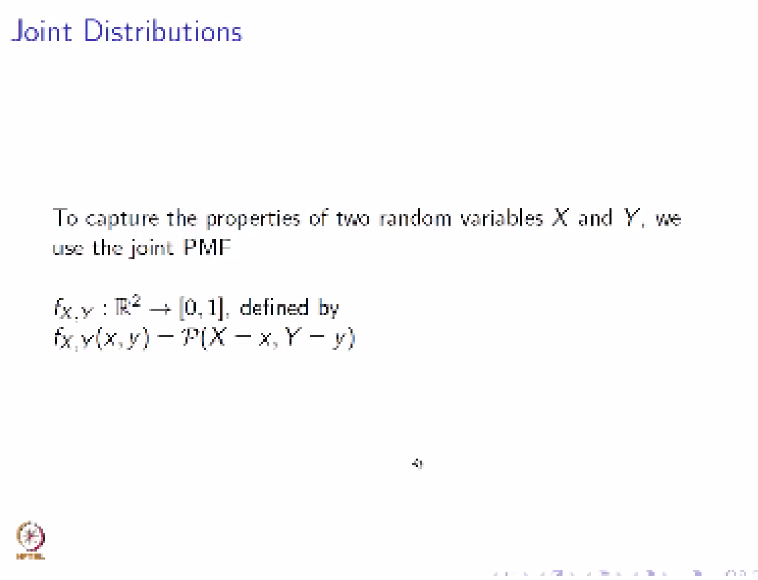
Closely related to the concept of covariance is the concept of correlation the correlation of two random variables X and Y is nothing but the covariance of the two random variables X and Y divided by the square root of the of the product of their individual variances basically correlation is a normalized version of covariance, so the correlation will always be between - 1 and 1 also since we used the variance of the individual random variables in the denominator for correlation to be divine individual variances must be nonzero and finite.

(Refer Slide Time: 16:42)



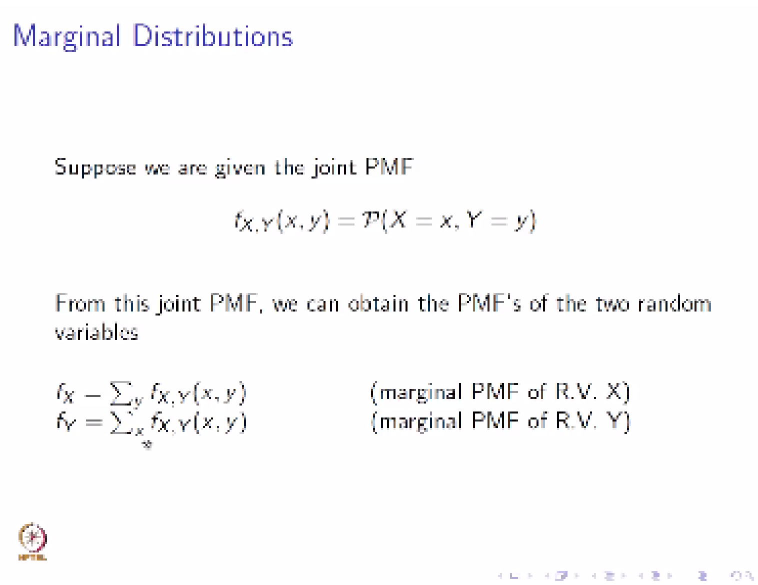
In the final part of this tutorial on probability theory we will talk about probability distributions and list out some of the more common distribution that you are going to encounter in the course before we proceed let us consider this question consider two variables x and y and suppose we know the corresponding probability mass function FX and FY corresponding to the variables x and y can we answer the following question what is the probability that X takes a certain value small x and y takes a certain value small y think about this question. If you answered no then you're correct let us see why.

(Refer Slide Time: 17:25)



Essentially what we were looking for in the previous question was the Joint Distribution which captures the properties of both the random variables the individual PMS or PDFs in case the random variables are continuous capture the properties of the individual random variables only but miss out on how the two variables are related thus we define the joint PMF or PDF, FXY as the probability that X takes on a specific value small x and y takes on a specific value smaller y for all values of X & Y.

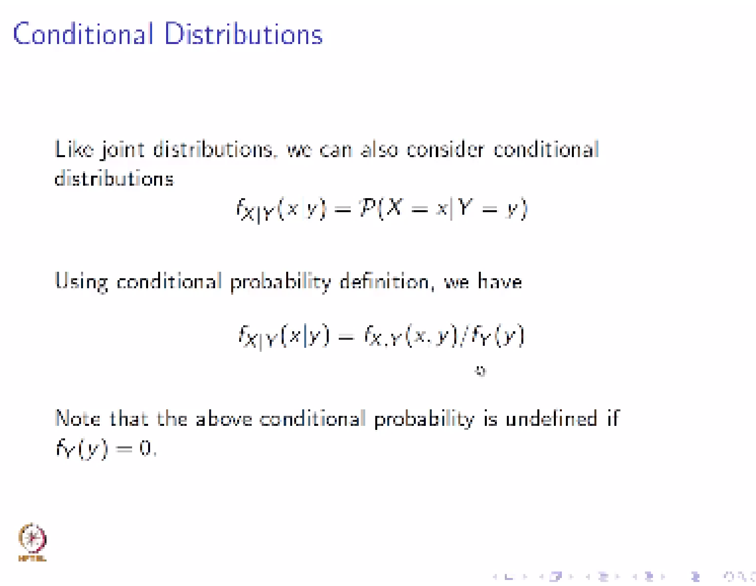
(Refer Slide Time: 18:02)



Suppose we are given the joint probability mass function of the random variables x and y what if we are interested in only the individual mass functions of either of the random variables this can be obtained from the joint probability mass function by a process called marginalization the individual probability mass function thus obtained is also referred to as a marginal probability mass function thus if you are interested in the marginal probability mass function of random variable x we can obtain this by summing the joint probability mass function overall values of Y.

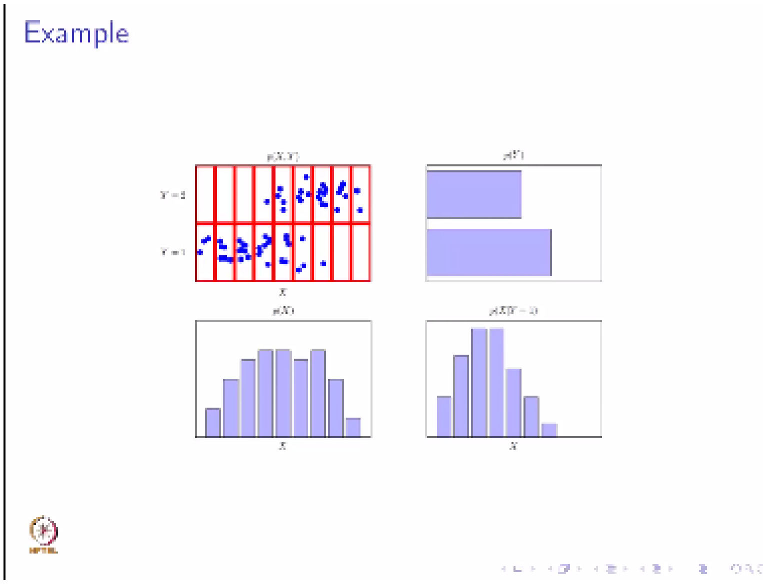
Similarly the probability mass function of the marginal property mass function of random variable Y can be obtained by summing the joint probability mass function over all values of x note that in case the random variables considered here are continuous we substitute summation by integration and PM/ PDF.

(Refer Slide Time: 19:02)



Like joint distributions we can also consider conditional distributions for example here we have the conditional distribution fx given Y which is the probability that the random variable X will take on some value small x given that the random variable Y has been observed to take on a specific value small y the relation between conditional distributions Joint Distribution and marginal distributions are is shown here this relation should be familiar from the definition of conditional probability that was seen earlier note that the marginal distribution fyy is in the denominator. And hence it must not be equal to 0.

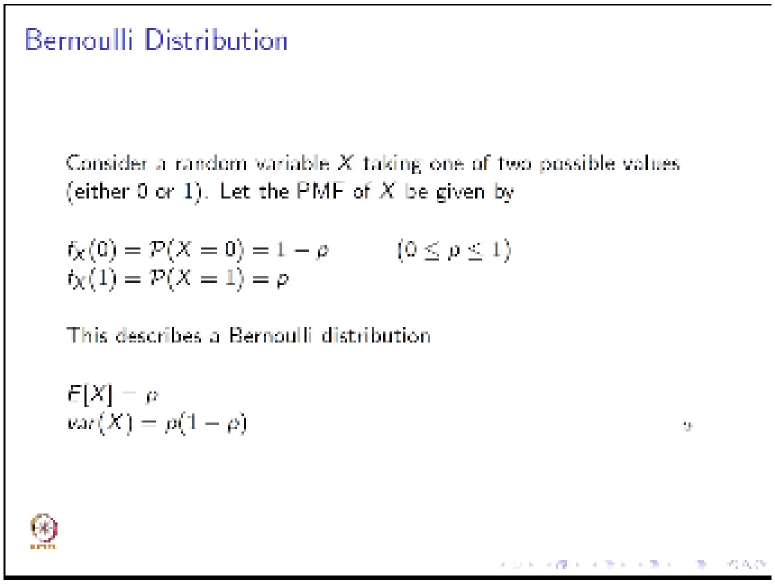
(Refer Slide Time: 19:45)



The overall idea of joint marginal and conditional distributions is summarized in this figure the top left figure shows the Joint Distribution and describes how the random variable X which takes on 9 different values is related to the random variable Y which takes on two different values the bottom left figure shows the marginal distribution of random variable X as can be observed in this figure we ignore the information related to the random variable Y.

Similarly the top-right figure shows the marginal distribution of random variable Y finally the bottom-right figure shows the conditional distribution of x given that the random variable Y takes on a value of one looking at this figure and comparing it with a joint distribution we observe that in the bottom-right figure is simply ignore all the values of x for which y equals to 2 that is the top half of the joint distribution.

(Refer Slide Time: 20:46)

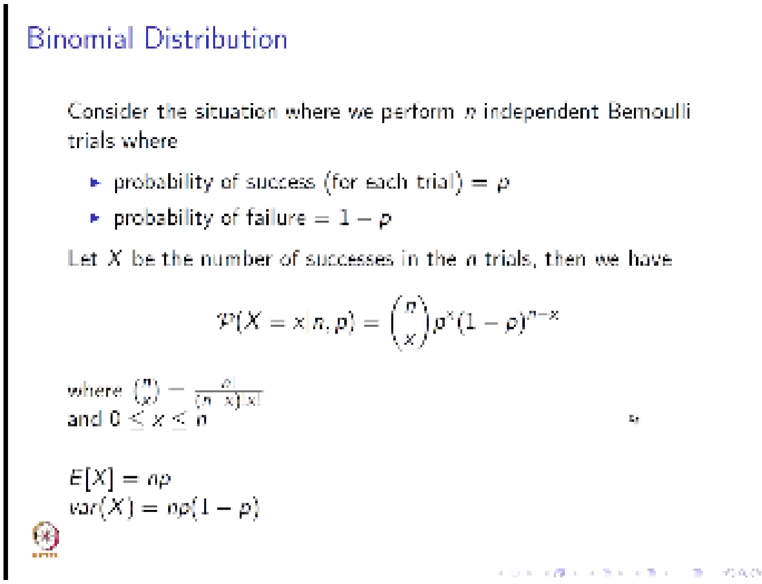


In the next few slides we will present some specific distributions that you will be encountering in the machine learning course we will present the definition and list out some important properties for each distribution it would be a good exercise for you to work out the expressions for the pms or PDFs and the expectation and variances of these distributions on your own we start with the Bernoulli distribution consider a random variable X taking one of two possible values either 0 or 1 let the PMF of X be given by FX of 0 is equal to probability that the random variable X takes on a value of 0 = 1-P where P is lies between 0 and 1and FX of 1 equals to probability that the random variable X takes a value 1 = p.

Here p is the parameter associated with the Bernoulli distribution it generally refers to the probability of success so in our definition we are assuming that X = 1 indicates a successful trial and x equals to 0 indicates of failure the expectation of a random variable following the Bernoulli distribution is P and the variance is P x 1 – P the Bernoulli distribution is very useful to characterize experiments which have a binary outcome such as in tossing a coin we observe either heads or tails or say in writing an exam either pass or fail.

Such experiments can be modeled using the Bernoulli distribution next we look at the binomial distribution.

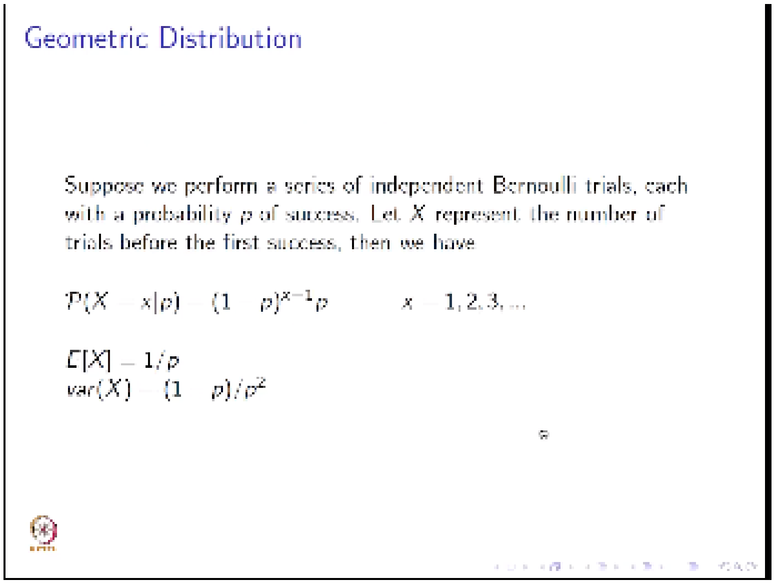
(Refer Slide Time: 22:18)



Consider the situation where we perform n independent Bernoulli trials where the probability of success for each tribe equals to P and the probability of failure for each trial equals to 1 - B Let X be the random variable which represents the number of successes in the end trials then we have probability that the random variable X will take on a specific value of small x given the parameters n and P = n choose X that is the number of combinations of observing X successes in n trials in to b^x x 1 – p^{n-x} .

Note that here x is going to be a number between 0and n the expectation of a random variable following the binomial distribution equals to NP and the variance equals to n x P x 1 - P the binomial distribution is useful in any scenario where we are conducting multiple Bernoulli trials that is experiments in which the outcome is binary for example suppose we have a coin suppose we toss a coin 10 times and want to know the probability of observing three heads given the probability of observing a head in an individual trial we can apply the binomial distribution to find out the required probability.

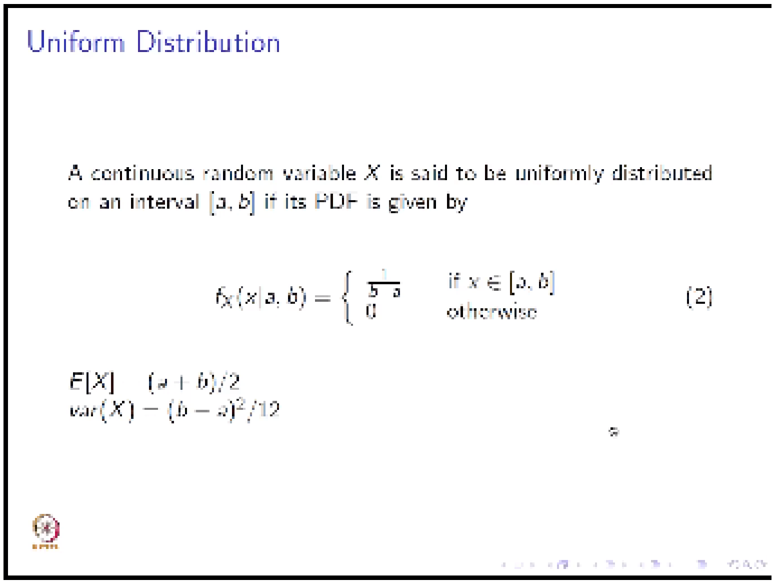
(Refer Slide Time: 23:38)



Suppose we perform a series of independent Bernoulli trials each with the probability P of success let X represent the number of trials before the first success then we have probability that the random variable X will take a value small x given the parameter P = 1 – p^{x} - 1 x p this definition is quite intuitive. Essentially we are trying to calculate the probability that it takes us small X number of trials before observing the first success this can happen if the first x minus y x minus 1 trials failed that is with probability 1 - P and the last I will succeeded that is with probability P.

A random variable which has the this probability mass function follows the geometric distribution for the geometric distribution the expectation of the random variable equals to 1 / P and the variance equals to 1 - P / P^2.

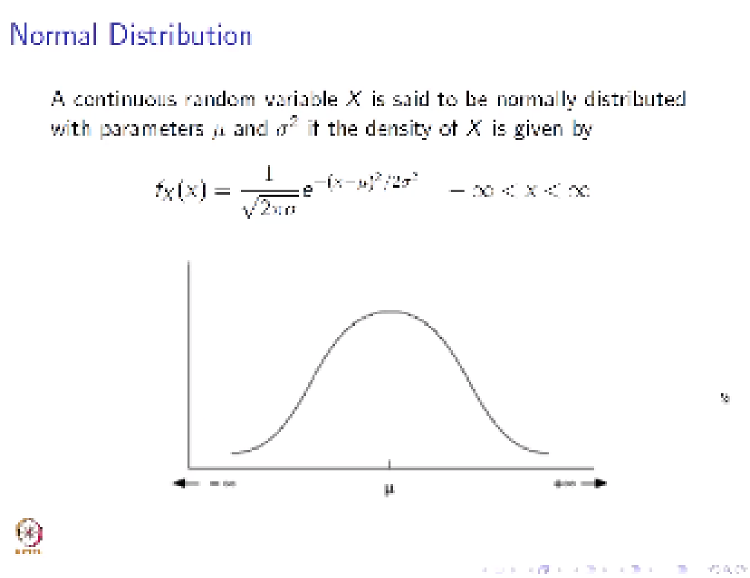
(Refer Slide Time: 24:38)



Many situations we initially do not know the probability distribution of the random variable under consideration but can perform experiments which will gradually reveal the nature of the distribution in such a scenario we can use the uniform distribution to assign uniform probabilities to all values of the random variable which are then later up dated in the discrete case say the random variable can take n different values then we simply assign a probability of 1 / n to each of the N values in the continuous case if the random variable X takes values in the closed interval a comma B then it is PDF is given by FX of x given parameters a comma B = 1 / B - A if X lies in the end closed interval a comma B and 0 otherwise.

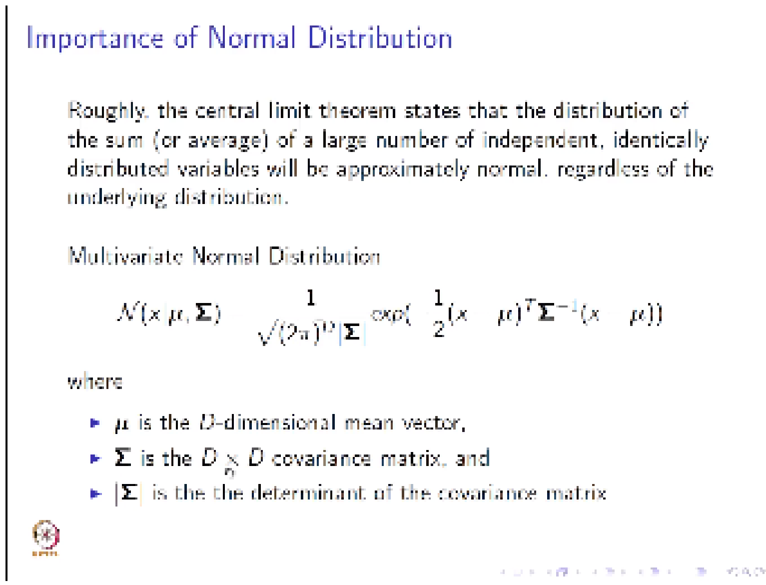
For a random variable following the uniform distribution the expectation of the random variable x = 2a + B / 2 and the variance equal to B – a^2 the B – A^2 / 12.

(Refer Slide Time: 25:48)



A continuous random variable X is said to be normally distributed with parameters mu and Sigma square if the PDF of the random variable X is given by the following expression the normal distribution is also known as the Gaussian distribution and is one of the most important distributions that we will be using the diagram represents the famous bell-shaped curve associated with the normal distribution.

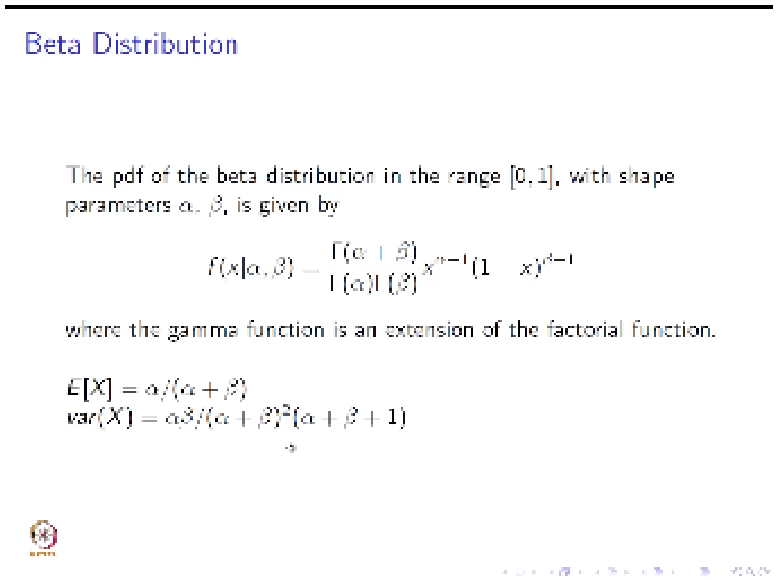
(Refer Slide Time: 26:15)



The importance of the normal distribution is due to the central limit theorem without going into the details there are central limit theorem roughly states that the distribution of the sum of a large number of independent identically distributed variables will be approximately normal regardless of the underlying distribution due to this theorem many physical quantities that are the sum of many independent processes often have distributions that can be modeled using the normal distribution.

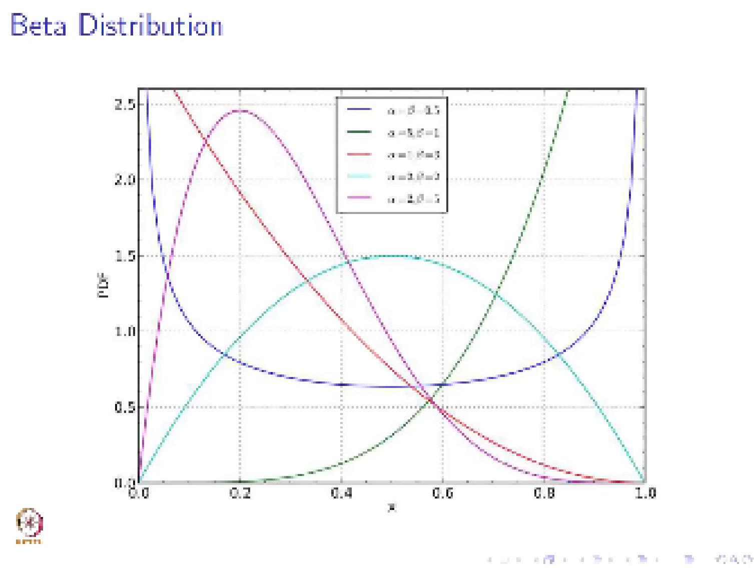
Also in the machine learning course we will be often using the normal distribution in its multivariate form here we have presented the expression of the multivariate normal distribution where μ is the D dimensional mean vector and Σ is the D cross D covariance matrix.

(Refer Slide Time: 27:06)



The PDF of the β distribution in the range 0 to 1 which shape parameters α and β is given by the following expression where the λ function is an extension of the factorial function the expectation of a random variable following the β distribution is given by α /α + β and the variance is given by α β / α + β^2 x α + β + 1.

(Refer Slide Time: 27:32)



This diagram illustrates the β distribution similar to the normal distribution in which the shape and position of the bell curve is controlled by the parameters μ and σ^2 in the β distribution the shape of the distribution is controlled by the parameters α and β in the diagram we can see a few instances of the β distribution for different values of the shape parameters note that unlike the normal distribution a random variable following the beta distribution takes values only in a fixed interval.

Thus in this example the probability that the random variable takes a value less than 0 or greater than 1 = 0 this ends the first tutorial on the basics of probability theory, if you have any doubts or seek clarifications regarding them regarding the material covered in this tutorial please make use of the forum to ask questions as mentioned in the beginning if you are not comfortable with any of the concepts presented here do go back and read up on it there will be some questions from probability theory in the first assignment. So hopefully going through this tutorial will help you in answering those questions and note that there we will be having another tutorial next week on linear algebra.

IIT Madras Production

Funded by
Department of Higher Education
Ministry of Human Resource Development
Government of India

www.nptel.ac.in