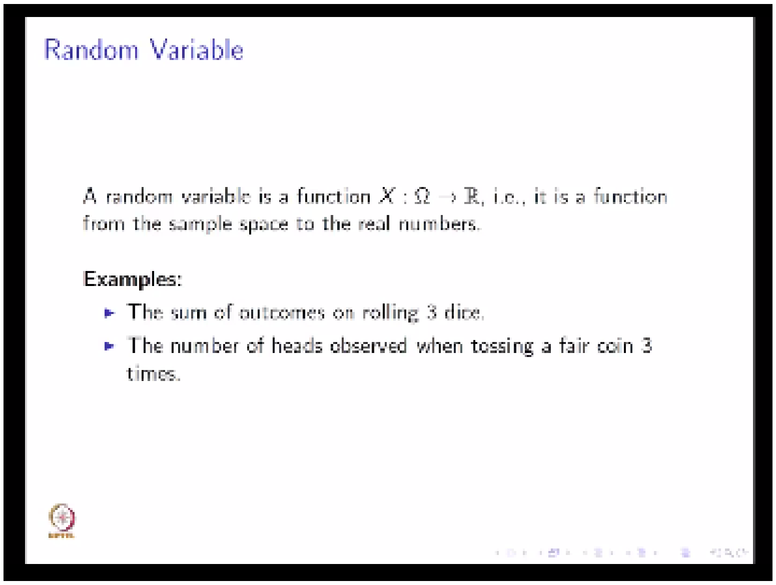
NPTEL

NPTEL ONLINE CERTIFICATION COURSE

Introduction to Machine Learning

Probability Basic-2

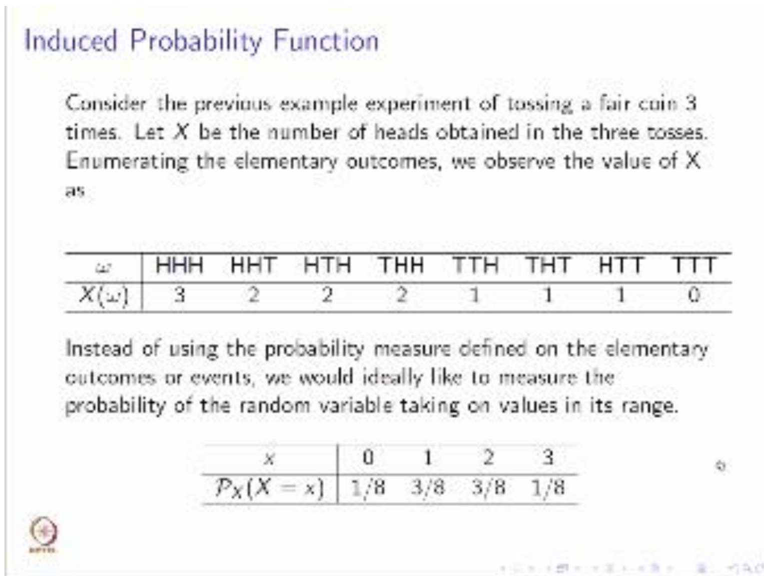
(Refer Slide Time: 00:15)



One of the important concepts in probability theory is that of the random variable our random variable is a variable whose value is subject to variations that is a random variable take on set of possible different values each with an associated probability mathematically a random 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 3 dice and are interested in the sum of the outcomes for example the sum of 5 can be observed if two of the dice show up two piece and the other dice shows up as 1.

Alternatively the sum of 5 can also be observed if one shows up as 3 and the other 2 dice show up 1 each since we are interested in only the sum and not the individual results of the dice rolls we can defined a random variables which maps the element way outcomes that is the outcomes of each die rolled to the sum up the three rolls similar in the next example we can defined a random variable which counts the number of heads observed when tossing a fair coin 3 times note that in this example the random variable can take values between 0 and 3 where as in the previous example the range of random variable is between 3 and 18 corresponding to all dice showing up 1 and all dice showing up 6.

(Refer Slide Time: 01:44)

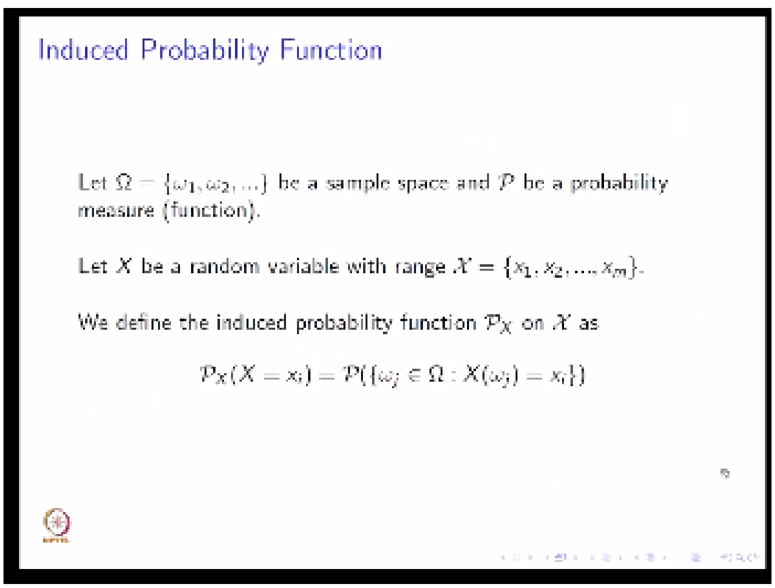


Consider the previous example experiment of tossing of fair coin three times let x be the number of heads obtained in the 3 tosses that is x is a random variable which maps each elementary out come to a real number representing the number of the heads observed in that outcome this is shown in the first table the first row list out each elementary outcome and the second row list 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 it is range what we are trying to say here is that when we defined probability measure we where associating each event that is sub set of the sample space with the probability measure when we consider random variables the events correspond to deferent subsets of the sample space which map two different values of the random variable.

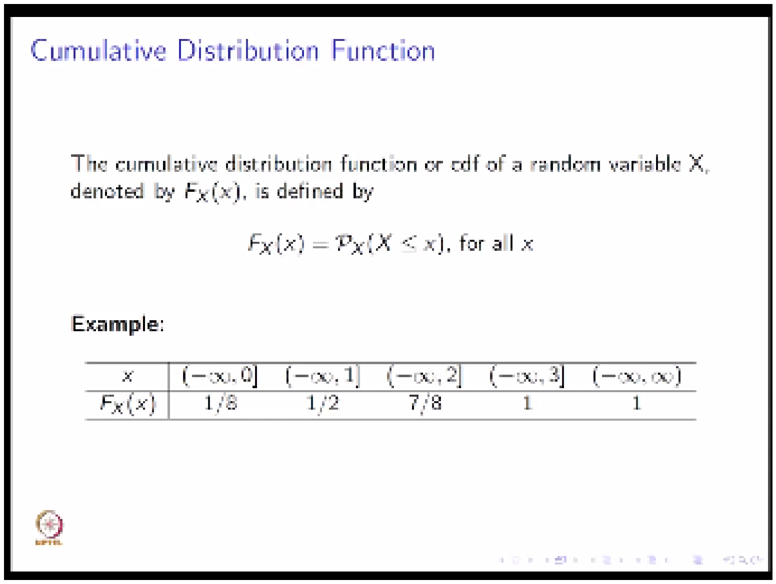
This is illustrated in the second table the first row list out the different value that the random variable X can take and the second row list out the corresponding probability values assuming that coin tossed is a fair coin this table describers an motion of the induced probability function which maps each possible value of the random variable to it is Associated probability value for example in the table the probability of the random variable taking the value of 1 is given as 3/ 8 since there are 3 elementary outcomes in which only one head is observed and each of these elementary outcomes as probability of 1/ 8.

(Refer Slide Time: 03:29)



From the previous example we can define the concept of the induced probability function let Ω be a sample space and P be probability measure let x be random variable which takes values in the range X_1 TO X_M the induced probability function Px on X is defined as Px X =xi = to the probability of the event comprising of the elementary outcomes ωj such that the random variable x mat ωj to the value xi.

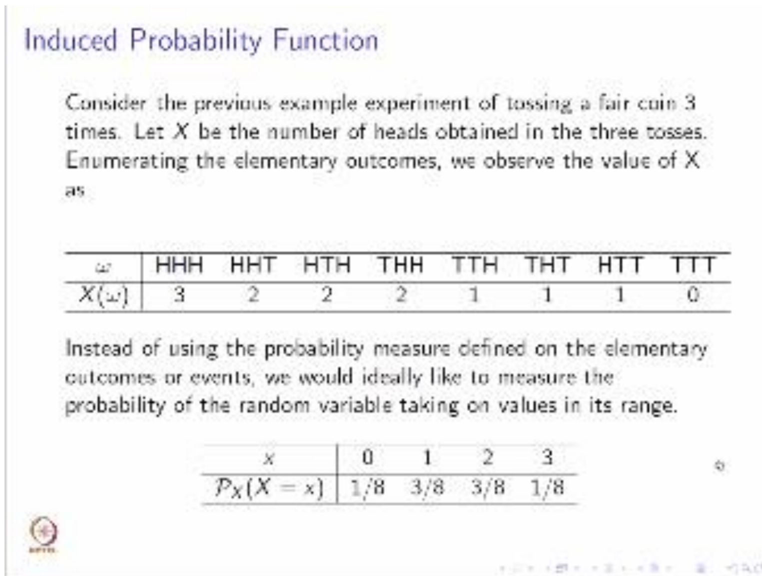
(Refer Slide Time: 04:11)



The cumulative distribution function or cdf our random variable X denoted by $F_X(x)$ is defined by $F_X(x)$ -= the probability of the random variable taking on a value less than or equal to x for all value of x for example going back to the previous random variable which counts a number of head observed in 3 toss of a fair coin the following table shows the intervals corresponding to the different values of the random variable x along with the corresponding value of the cumulative distribution function.

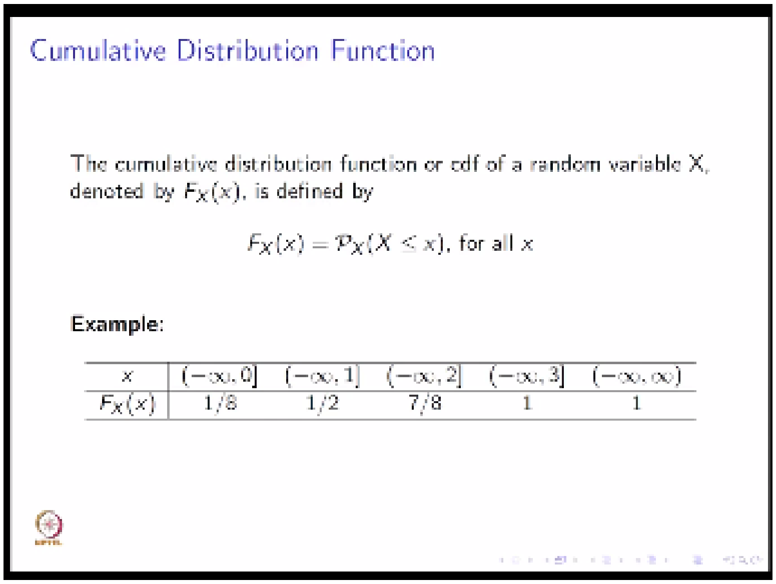
For example $F_{X=}$ $F_X(1)$ = ½ because the probability that the random variable X as a value of 1 now lets us go back to the previous example.

(Refer Slide Time: 05:04)



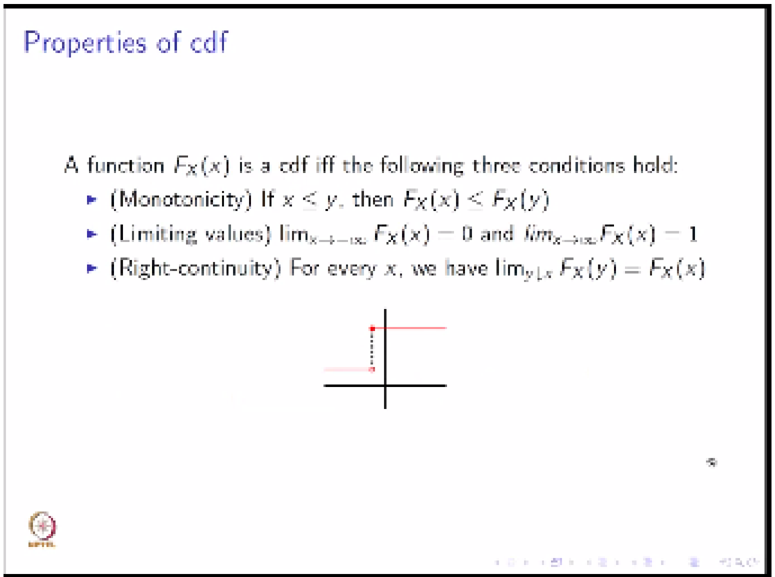
Right the probability that the random variable x as a value 1 is 3/8 the probability of x that the random variable x = 0 is 1/8 and therefore.

(Refer Slide Time: 05:15)



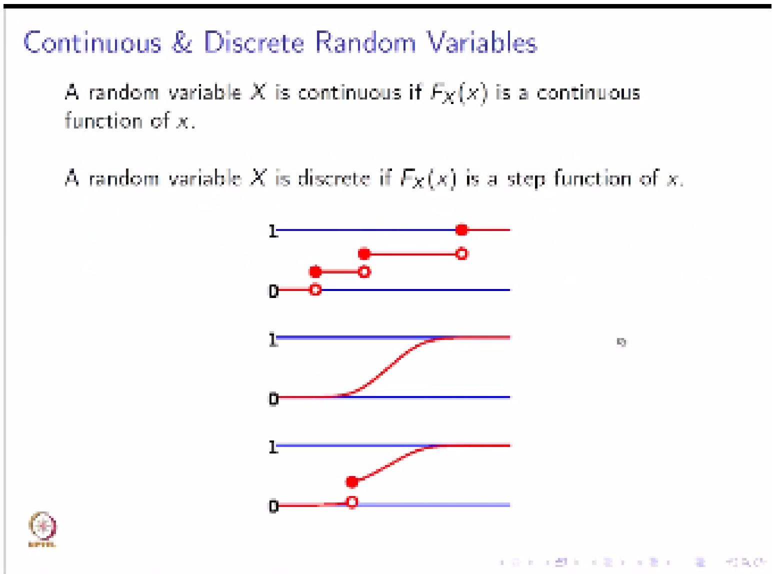
The probability that the random variable x takes on value less than or equal to 1 is 1/8+ 3/8 = 4/8 or ½

(Refer Slide Time: 05:26)



a function is a valid cumulative distribution function only if it satisfies the following properties the property simply states that the cumulative distribution function is non decreasing function the second properties specifies a limiting values limit extents to - ∞ Fx(x) = 0 and limit extends to ∞ Fx(x) = 1 the third property specifies right continuity that is now jump occurs when the limit point is approach from the right this is also shown in the figure below.

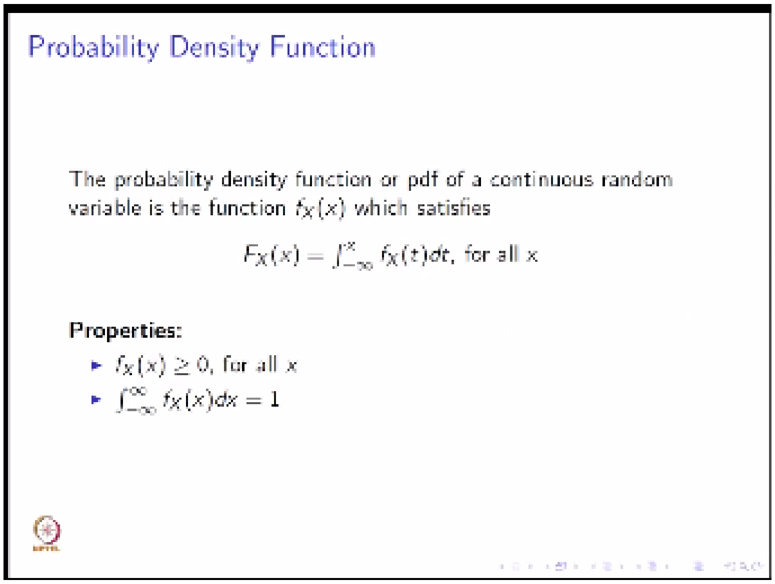
(Refer Slide Time: 06:04)



A random variable x is continuous if it is corresponding cumulative distribution function is a continuous function of X this is shown in the second part of the diagram random variable X is discrete if it is cdf irrespective function F this is shown in the first part of the diagram the third part of the diagram shows the cumulative distribution function or a random variable which as those continuous and discrete parts.

\

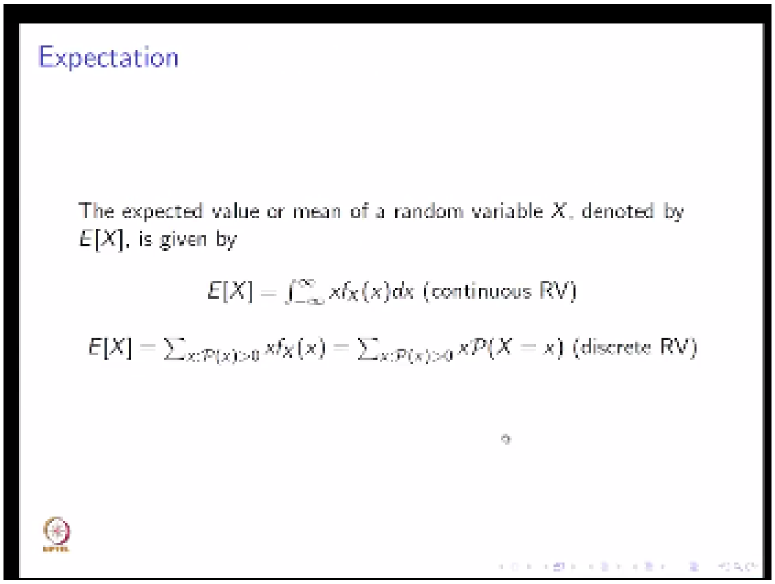
(Refer Slide Time: 06:33)



The probability mass function of pmf of a discrete random variable X is given by Fx(x) = probability of X = x or all values are x thus for a discrete random variable the probability mass function of that random variable gives the probability that the random variable is equals to some value for example for a geometric random valuable X with parameter P the PMF is given as Fx(x) = 1-P $^{x-1}$ v 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 properties first of all the function must be non negative secondly the Σ over all X the value of the function summed over all values of X should be =1.

For continuous random variables we consider the probability density function or PDF of a ran continuous random variable is the function Fx(x) which satisfies the following the integral form -∞ to x Fx(tdt) is = 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 value of X second integrating over the entire range the probability density function should sum to 1.

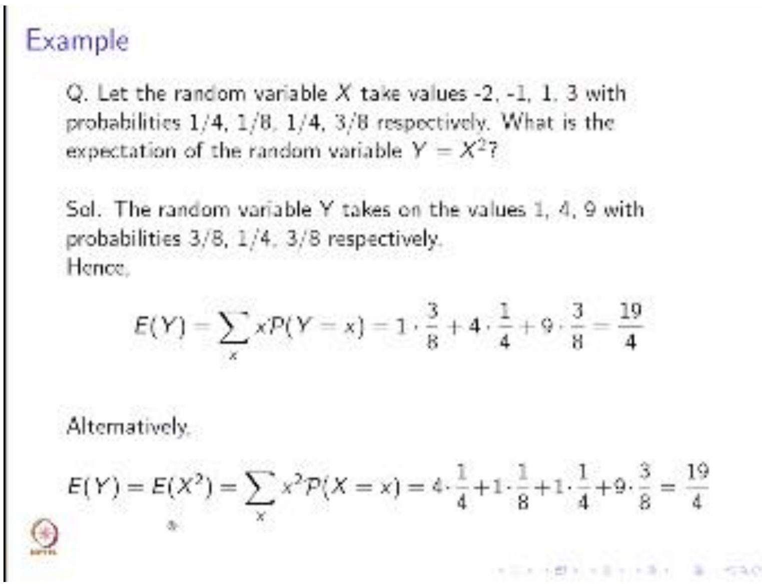
(Refer Slide Time: 08:20)



Let us now look at expectations of random variables the expected value for mean of a random variable x denoted by a expectation of x is given by integral - ∞ to ∞ x into Fx(x) dx now that Fx)x) is here is the probability density function associated with the random variable X this detention holds when x is a continuous a random variable in case that X is a discrete random variable we use the fowling definition expectation of x is = 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 0 x x Fx(x).

Here Fx(x) is the probability mass function of the random variable x which essentially give 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 1/4, 1/8. ¼ and 3/8 respectively what is the expectation of a random variable $Y = X^2$ so in this question we are given one random variable the value which this random variables takes and it is associated probabilities what we are interested in the expectation of the random variable Y which is defined as $Y= X^2$ so what we can do is we can calculate the values that the random variable Y take along with the associated probabilities since we are aware of the relation between Y and X.

Thus we have Y taking on the value 1, 4 and 9 with probability 3/8, 4/4 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 give a result 19/4 another way to approach this problem was it is to directly use the relation $Y= X^2$ in calculating the expectation thus expectation 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 sum over all X X^2 into probability of X= X calculating the values we keep the same answer of 19/4.

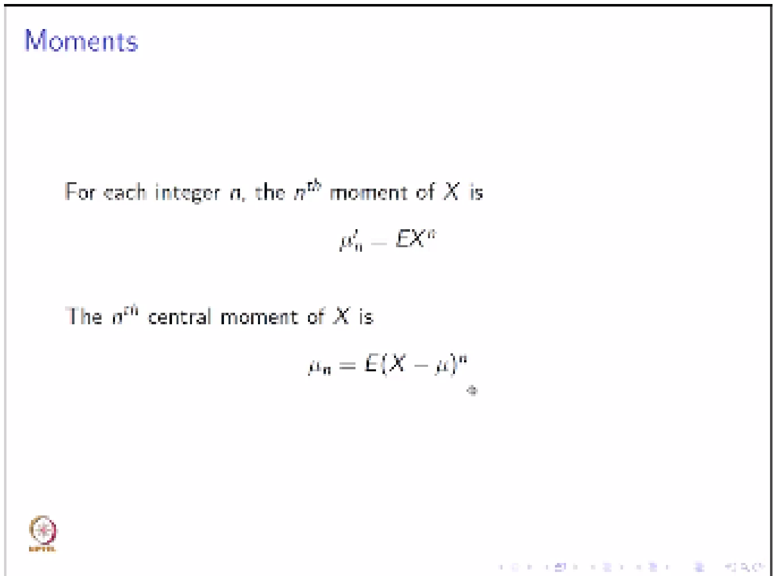
(Refer Slide Time: 10:55)

Let us now look at the properties of expectations let X be a random variable AB and C constants and G1 A and G2 are function of the random variable X such that they are expectations exits that is they have finite expectations according to the first property expectation of Ax G1(x) + b times G2(x) + C = A times expectation of G1(x) + B times expectation of G2 (x) + C this is called the linearity of the expectations there are actually a few things to note here first of all expectation of a constant is = to the constant itself.

Expectation constant times a random variable is equal to the constant into the expectation of a random variable and the expectation of the sum of two expectations can also be represented as the sum of the expectations of the two random variables note that here the two random variable need not be strategically independent according to the next property if a random variable is greater than equals to 0 at all points then the expectation is also expectation of that random variables is also greater than equals to 0.

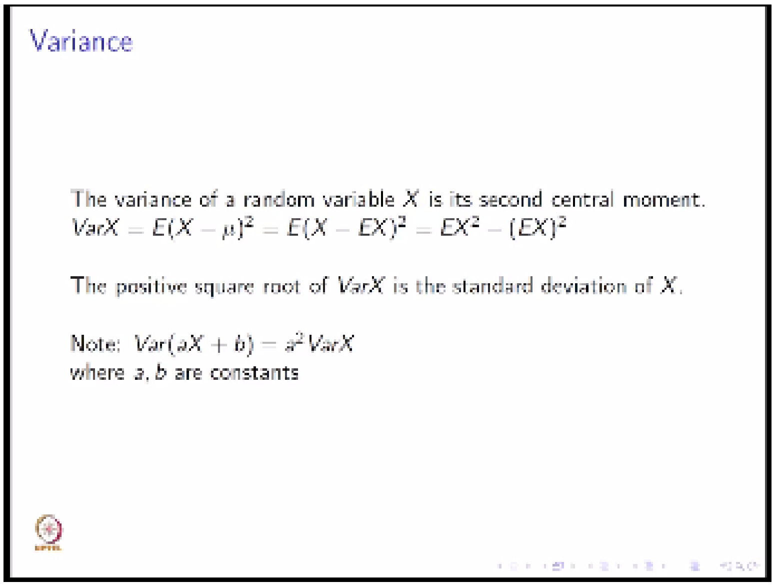
Similarly if one random variable is greater than another random variable at all points then the expectation of those random variables also follow the same constrained finally if a random variable as values which are which lie between two constants then the expectation of that random variable will also lie between those two constants.

(Refer Slide Time: 12:35)



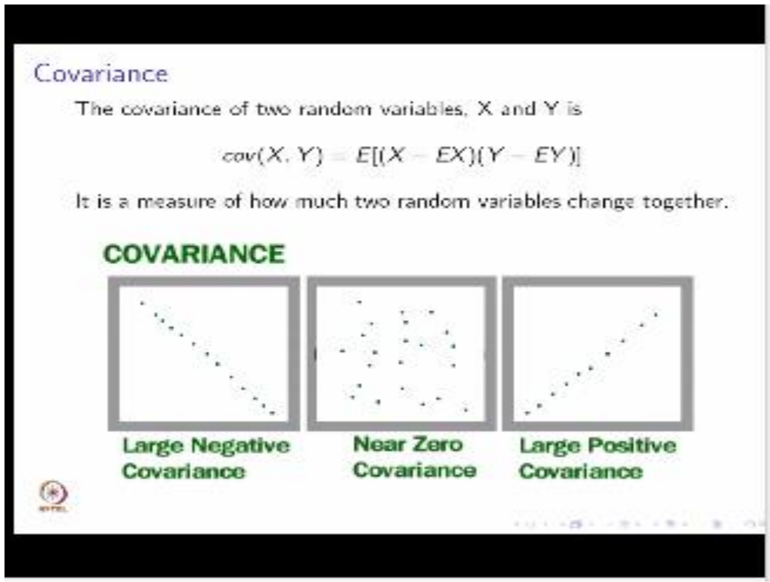
Let us now defined movement for integer n the n^{th} movement of X is μ' or n = $E(X)^n$ also the n^{th} central movement of X μn – $E(x- \mu)^n$ so the difference between movement and central movement is in central movement we subtract the random variable by the mean of the random variable or E value the two movements that 5 most common use are the first movement which is nothings but μ' = E(x) that is the mean of the random variable X and the second central movement which is μ2= $E(x- \mu)^2$ which is the variance of the random variable X.

(Refer Slide Time: 13:22)



Thus the variance of the random variable X is it second central movement variance of X = $E(x-\mu)^2$ now that μ is just the first movement which can be replaced so it can be replaced by E(x) thus we have variance of X = $E(X- EX)^2$ by expanding this term and applying linearity expectations we will finally get variance of X = EX2- X2 of the EX the positive square root of variance of X is a standard deviation of X note that the when as calculating variance the constants add differently and compared to the linearity expression, this is a very useful relation to remember, variance of (aX+b) = a^2Var(X), where a and b are constants.

(Refer Slide Time: 14:16)



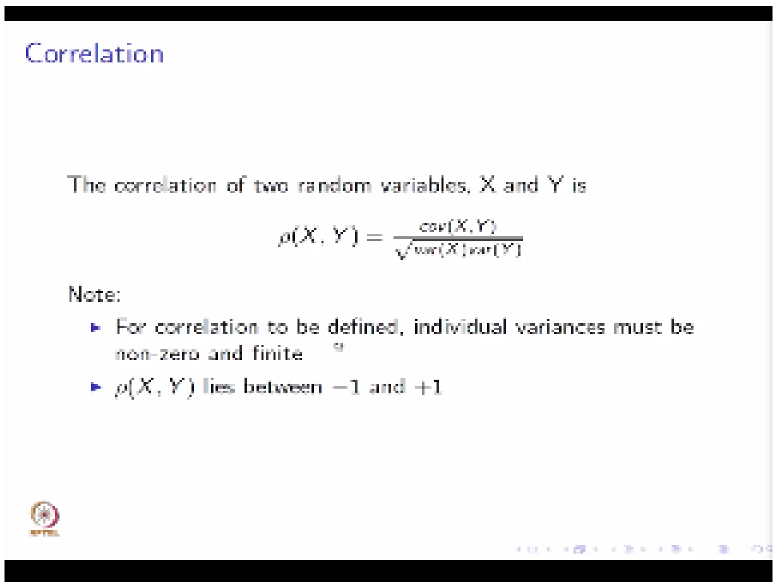
The co variance of two random variables x andy is, cov(X,Y)=E[{X EX}{Y EY}], remember that the variance, however the random variable X is nothing but the second central movement, thus the variance of a random variable, measures the amount of separation in the b values of the random variable, when compare to the mean of the random variable.

For co variance the calculation is done, on a pair of random variables, and it measures how much 2 random variables change together, consider the diagram below, in the first part assume that the random variable is on the X- axis, and the random variable is on the Y-axis, we note that as the value of X increases, the value of Y seems to be decreasing, thus for this relationship, we will observe a large negative co variance.

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 larger positive co variance, however in the middle diagram, we cannot make any such statement, because as X is increases, there is no clear relationship as to how Y changes.

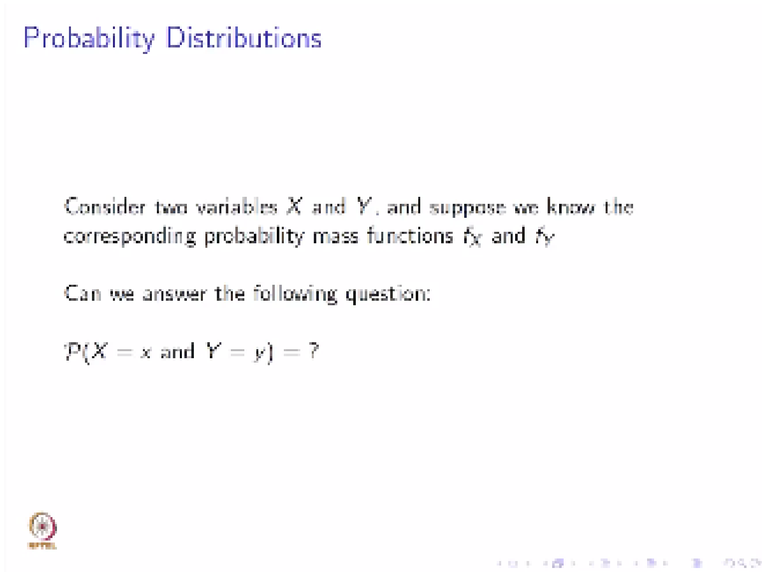
Thus this kind of a relationship will give 0 co variance, now from this diagram, it should immediately we clear that co variance is very important term in machine learning, because we are often interested, in predicting the value of 1 variable, by looking at the value of the other variable, we will come to that in further classes.

(Refer Slide Time: 16:04)



Closely related to the concept of the covariance is the concept of correlation, the correlation two random variables X and Y is nothing but the co variance of the two random variables p(X,Y)=cov(X,T)/ √var(X) var(Y), basically correlation is, normalized version of co variance, so the correlation will always between -1 and 1, also since we use the variance of the individual random variables, in the denominator the correlation to de divide , individual variances must be non 0 and finite.

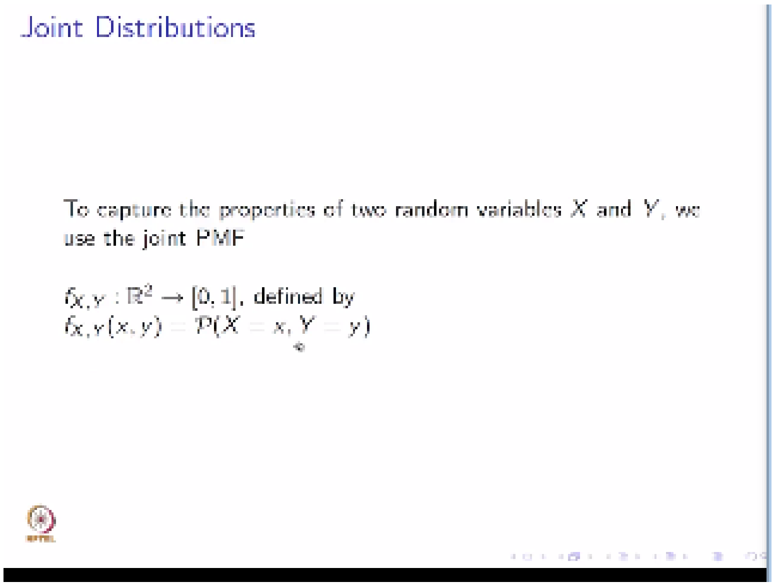
(Refer Slide Time: 16:42)



In the final part of this tutorial 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's consider this question, consider to variables, X and Y and suppose we know the corresponding probability mass function f_x and f_y corresponding to the variables action Y.

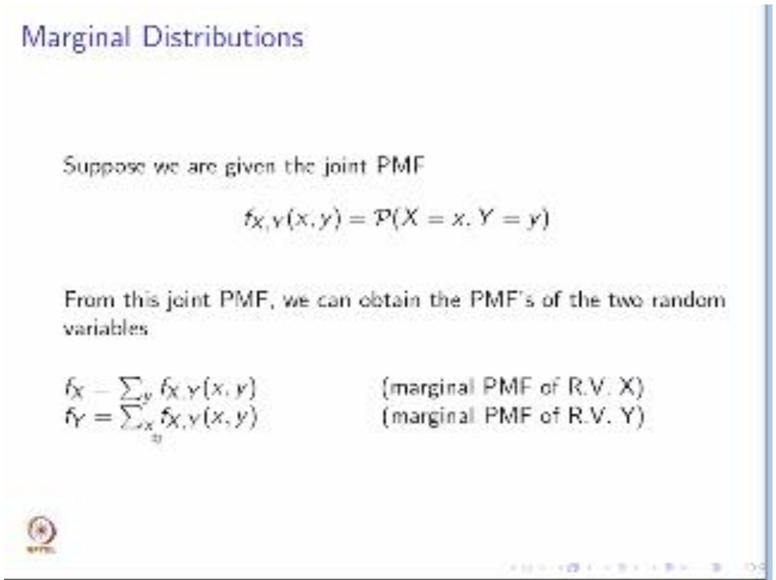
Can you answer the following question? What is the probability that X take a certain value small x, and Y take a certain value y, think about this question, if you answered no then you are correct, let see Y essentially what we are looking for the previous question, what the joint distribution which captures the properties of both the random variables. The individual PMF are PDF, in case the random variables are contentious.

(Refer Slide Time: 17:39)



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, $f_{x,y}$ as the probability that X takes on a specific value smaller, an Y takes on a specific value y, for all values of X and Y.

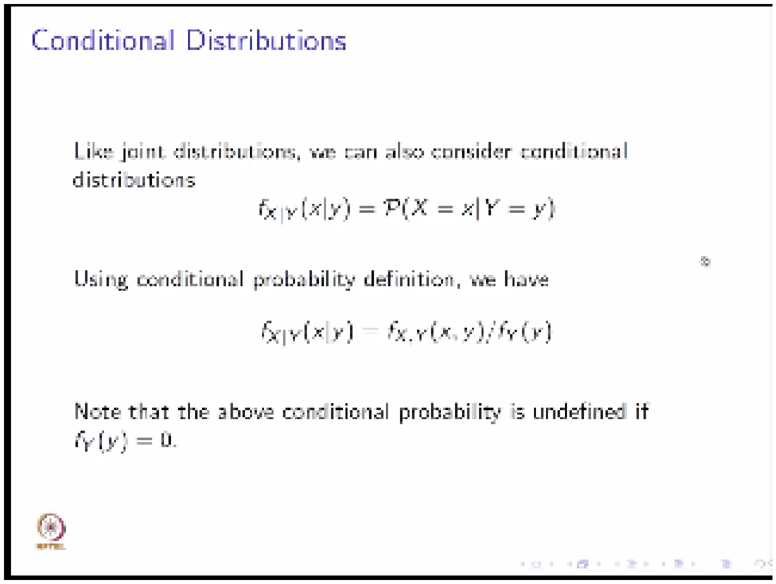
(Refer Slide Time: 18:02)



Suppose we are given the joint probability mass function of the random variables X and Y, what if you are interested 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 obtain is also refer thus the marginal probability mass function.

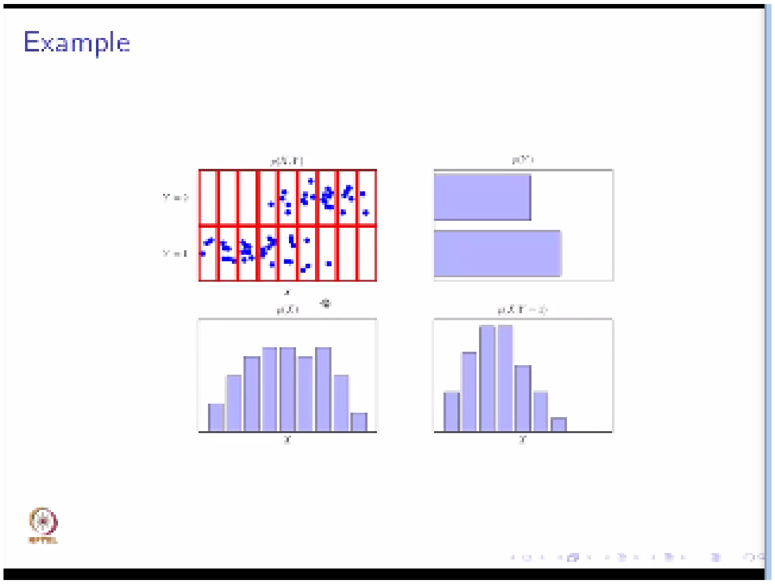
Thus if you interested in the marginal probability mass function of the random variable X, ewe can obtain this by summing the joint probability mass function over all values of Y, similarly the probability mass function of the marginal probability mass function of the random variable Y , can be obtain by summing the joint probability mass function over all values of X. Note that in case the random variables, considered here are contentious, we substitute summation by integration and PMF by PDF.

(Refer Slide Time: 19:01)



Like joint distribution we can also consider conditional distributions for example, here we have the conditional distribution $f_{xy}(X/Y)=P(X=x| \ Y=y)$, 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 we seen earlier. Note that the marginal distribution $F_y(Y)=0$ is in the denominator enhance it must no equals to.

(Refer Slide Time: 19:44)

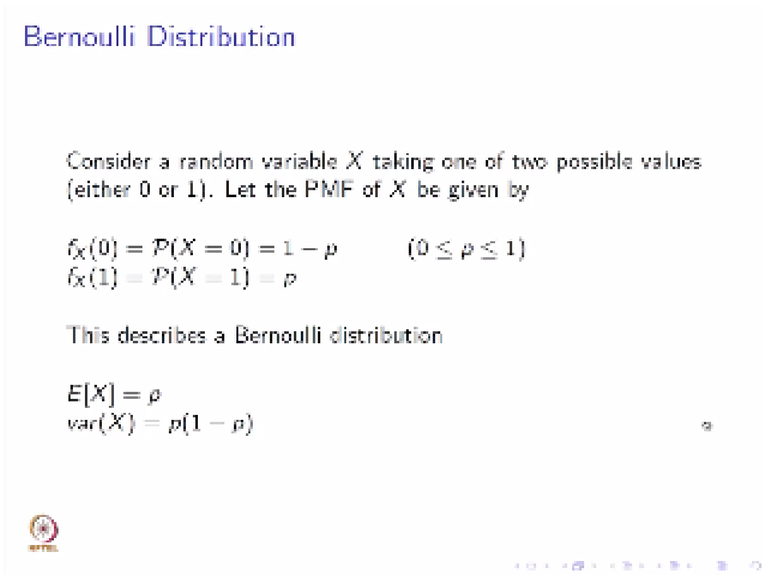


The overall idea of joint marginal and conditional distribution is summarized in this figure, the top left figure shows the joint distribution, and describe how the random variable X, which takes on the nine different values, is related to the random variable Y, which takes on two different values.

The bottom left figure shows the marginal distribution of random variables X, as can be observed in this figure we simply ignore the information to the random variable Y, similarly the top right figure shows the marginal distribution of the random variable Y, finally the bottom right figure shows the conditional distribution of X given the random variable Y, takes on a value of 1.

Looking at this figure and comparing it with the joint distribution, we absorbed that in the bottom right figure we simply ignore all the values of X for which Y=2, that is the top of their 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 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 PMF or PDF, and the expectations and the variances of these distributions on your own.

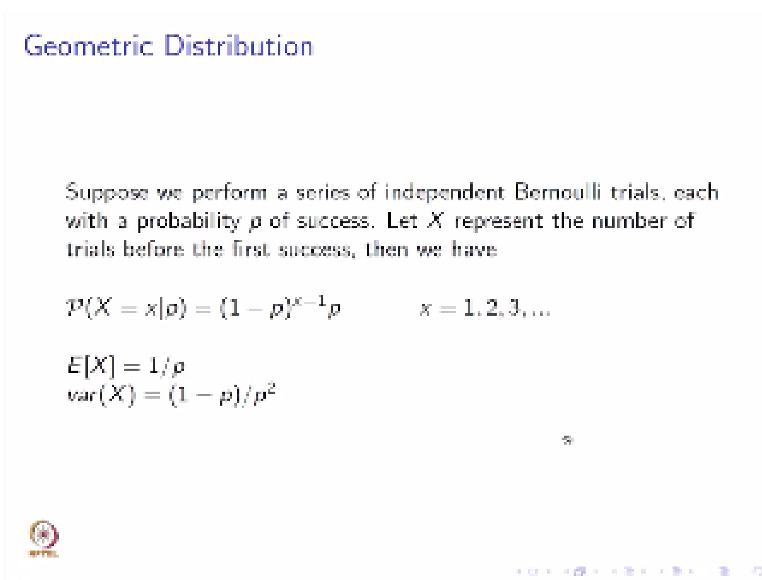
We start with the Bernoulli distribution , consider a random variable X taking one of two possible values, either 0 or 1, let the PMF of x given by $F_{X(0)}$=p(X=0) =1 where p lies between 0 and 1, and f_x(1)=p(X 1)p, here p is the parameter associated with the Bernoulli distribution.

It generally refers to the probability of success, so in our definition we are assuming that x=1 indicates the successful trial, and x=0 indicates the failure, the expectation of the random variable following the Bernoulli distribution is p, and the b variance is p (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 example these are pass or fail, such experiments can be modeled using the Bernoulli distribution.

Next we look at the binomial distribution, consider the situation when you perform an independent Bernoulli trials, where the probability of success for each trial equals to p, and the probability of earlier for each trial equals to 1-p, let x be the random variable, which represents the number of success in the N trials, then we have probability weather random variable X will take on a specific value of x, given the parameters nnp=n choose x that is the number of combinations observing x success in N trials into $p^x(1-p)^{n-x}$ here x is a number going to be 0 and n, the expectation of a random variable following the binomial distribution equals to np, the variance equals to np(1-p).

The binomial distribution is useful in any scenario where we are conducting multiple Bernoulli trials, that is experiments in this outcome is binary, for example suppose we have a coin, suppose we toss a coin 10 times, then want to know the probability of 3 heads, given the probability of observing an head n an individual trial we can apply the binomial distribution to find out the required probability.

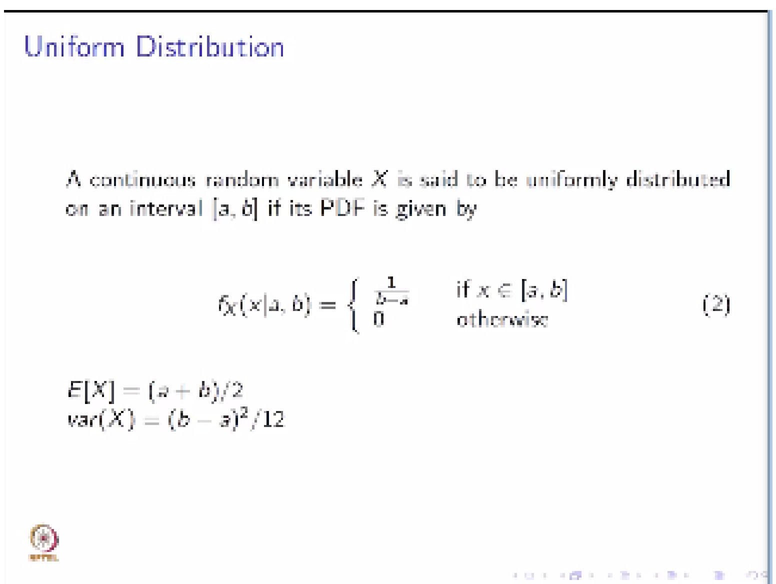
(Refer Slide Time: 23:38)



suppose we perform a series of independent Bernoulli trials, each with the probability each 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 Xx given the parameter P $=(1-p^{x-1})p$, this definition is quite inductive, essentially we are trying to calculate the probability that it takes us x number of trials before observing the first success.

This can happen if the first x-1 trials failed, that is with probability 1-p, when the last trial succeed, 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 equal to $1-p/p^2$

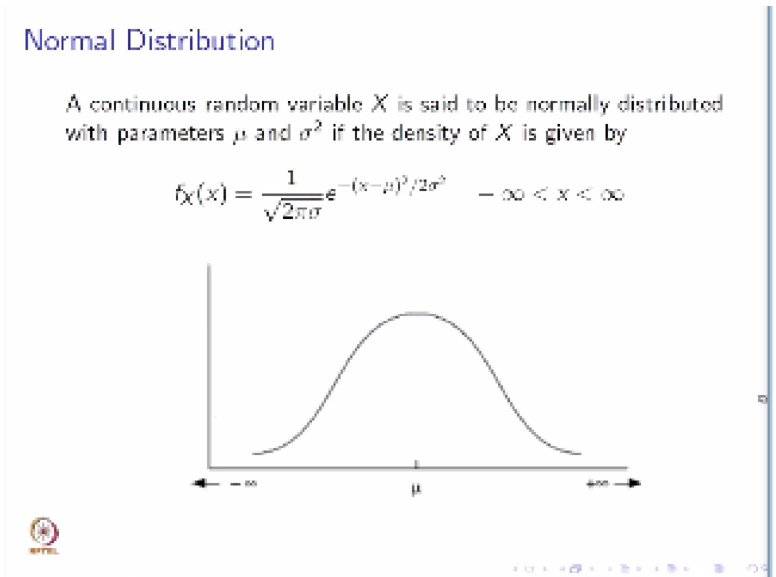
(Refer Slide Time: 24:40)



In many situation we initially do not know the probability distribution, of the random variable under consideration,, but can perform a experiments which will gradually revel 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 updated.

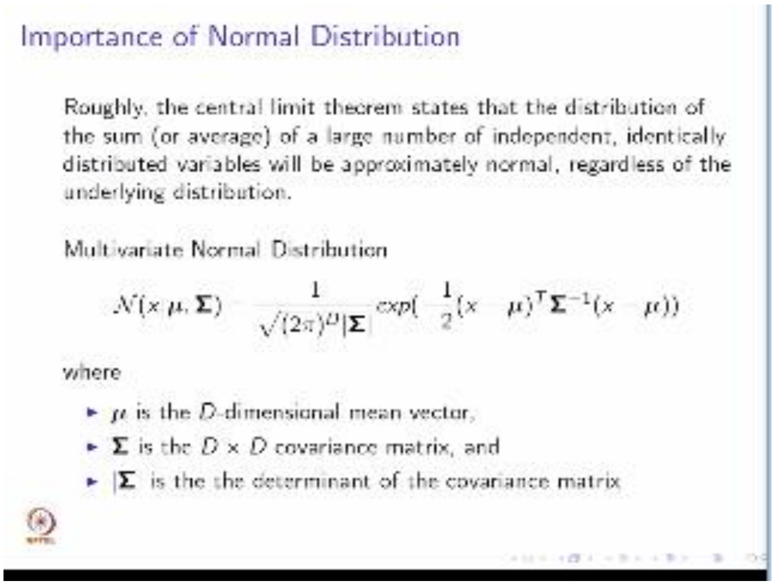
In the discrete case say the random variable can take N different values, the we simply assign a probability of 1/n to each of the N values, in the contentious case if the random variable X, takes values in the closed interval a,b then its PDF is given by $f_x(X[a,b])=1/b-a$ if x lies in the close interval a,b, and 0 otherwise. For a random variable following the uniform distribution the expectation of the random variable X=(a+b)/2, and the variance equals to $(b-a)^2/12$.

(Refer Slide Time: 25:49)



A continuous random variable X, is said to be normally distributed with parameters μ and σ^2, if the PDF at the random variable X is given by the following expression, its 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 shape curve associated with the normal distribution.

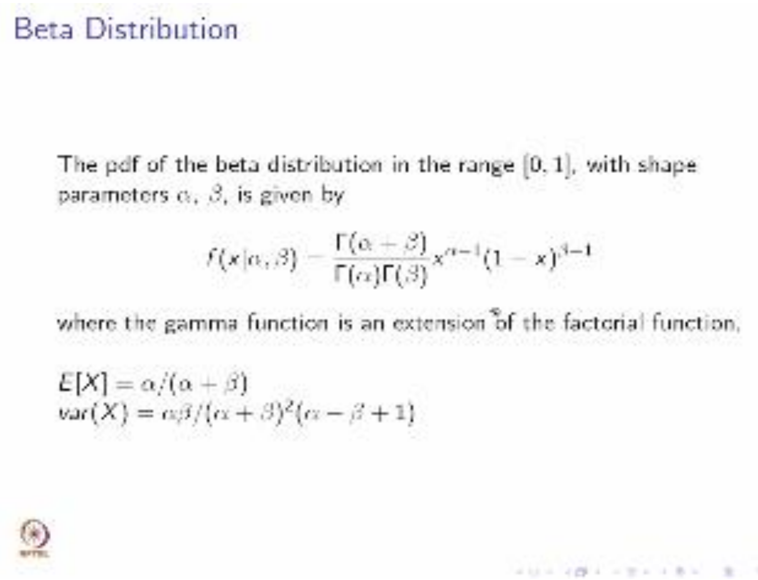
(Refer Slide Time: 26:16)



The importance of the normal distribution is due to the central limit theorem; without going into the details the central limit theorem roughly states the distribution of the some of the large number of independent identically distributed variables will be approximately normal, regardless of the underline distribution.

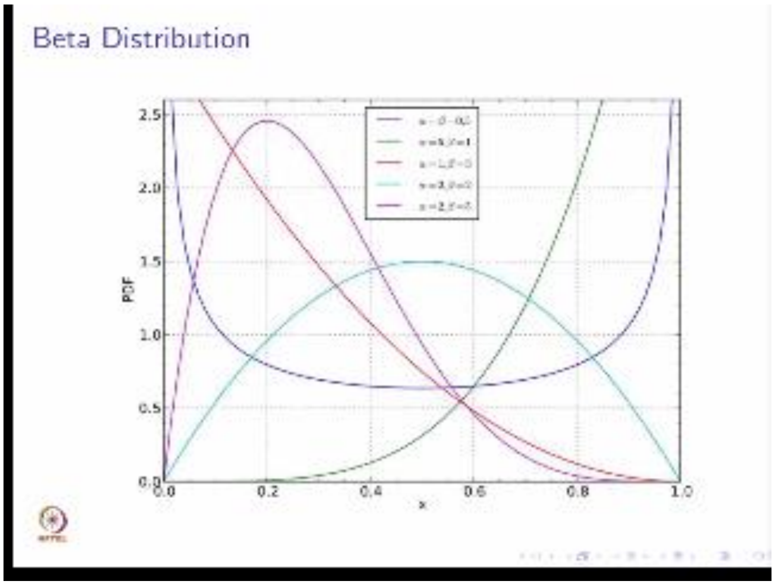
Due to this theorem many physical quantities, that are the sum of many independent process is often have distribution that can be modeled using the normal distribution, also in the machine learning course you will be often using the normal distribution in multivariate form, here we have presented the multivariate normal distribution, where μ is the D dimensional mean vector, and σ is the d×d covariance matrix. The PDF of the β distribution in the range 0 to 1, which says parameters α and β is given by the following expression.

(Refer Slide Time; 27:12)



Where the γ function is an extension of the factorial function, the expectation of the random variable following the β distribution is given by α / α +β, and the variance is given by α β/α +β^2 × α +β +1.

(Refer Slide Time; 27:33)



This diagram illustrates the β distribution similar to the normal distribution in which the shape and the position of the bell curve is controlled by the parameters μ and the σ^2 in the β distribution the shape of the distribution is controlled by the parameters α and β.

In the diagram e 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 for following the β variable distribution, only in a fixed intervals, this in this example probability the random variable takes the value less than 0 or greater then Is equals to 0.

This sends the first tutorial of basics of probability theory, if you have any doubts or clarification regarding the material covered in this tutorial please make use of the forum to ask the questions, has mention in the beginning, if you are not comfortable with the 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 to this tutorial will help you in answering those questions and note that we will 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