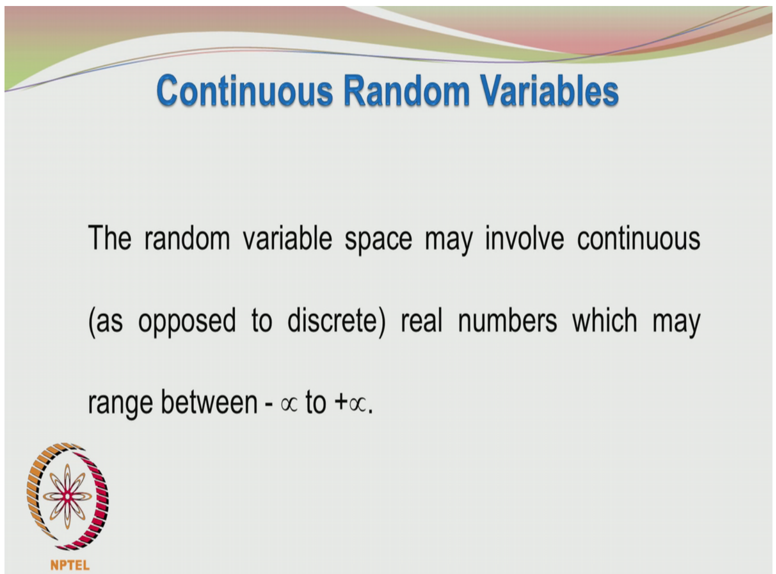
Statistics for Experimentalists
Prof. Kannan. A
Department of Chemical Engineering
Indian Institute of Technology - Madras

Lecture - 05
Continuous Probability Distributions

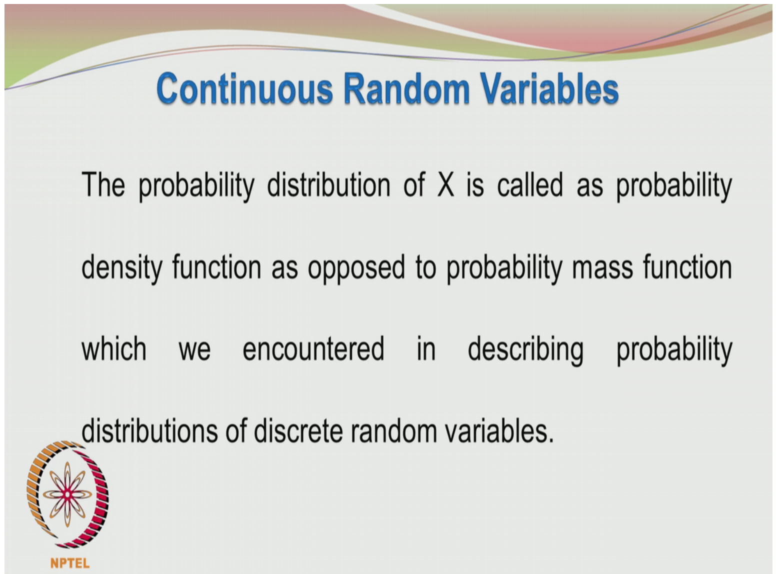
Welcome back, as I said earlier we will be now looking at Continuous Probability Density Functions.

(Refer Slide Time: 00:22)



The random variable x may either be discrete or continuous, it may take individual discrete values or it may take values within a continuous range okay, what I mean by a continuous range is even decimal numbers can be included. For example, mole fraction lies between 0 and 1, but it can take any value between 0 and 1.5, 3.5, 4 and so on. Whereas if you roll a die you can have only discrete entities like 1, 2, 3, 4, 5 and 6. So it may take real numbers which may range between -infinity to +infinity.

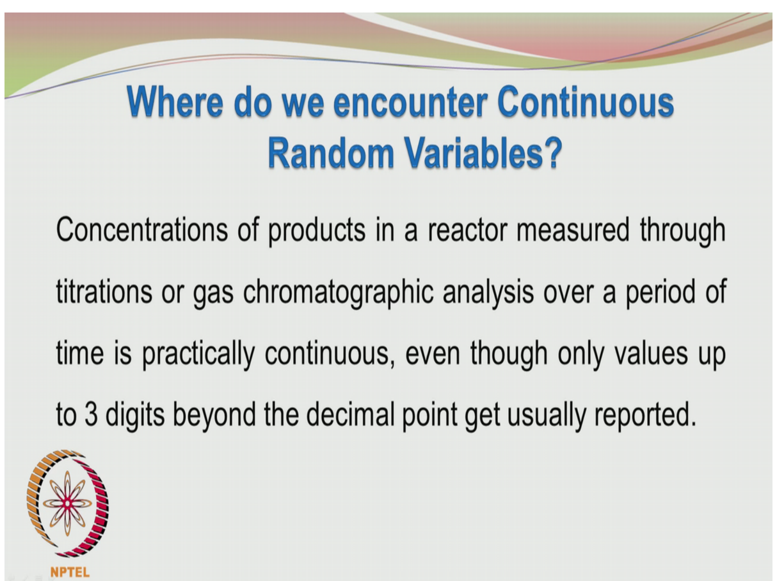
(Refer Slide Time: 01:16)



One of the important property of a random variable whether it is continuous or discrete is it has the probability distribution function associated with it, the probability distribution function simply means the probability associated with the random variable taking on a particular value or falling within a range of values. In the case of discrete probability distributions, the probability of a random variable taking a particular value may be given.

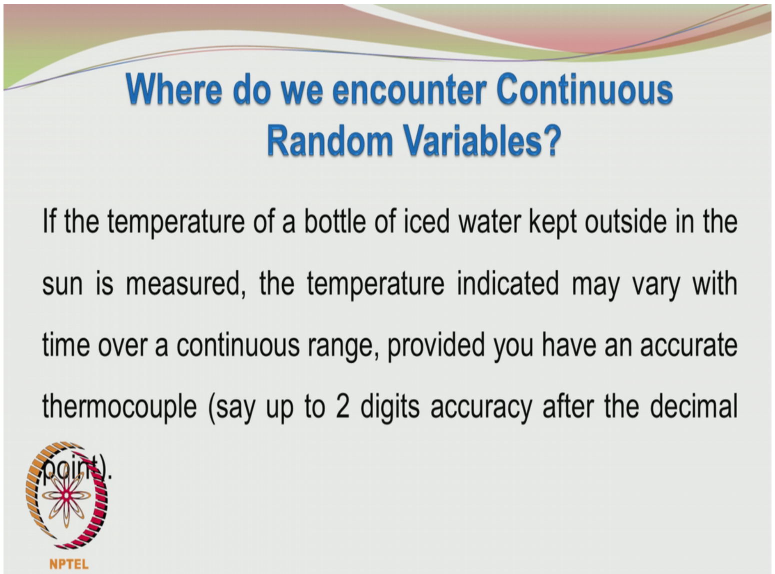
In the case of continuous random variables, you do not talk of the probability of a random variable taking a particular value, but you talk of a random variable taking a value between a certain range okay. We will be discussing more on this shortly.

(Refer Slide Time: 02:23)



Well now there is a chemical reaction taking place in a reactor, the products of the reaction are measured by using several analytical means, for example you may use a gas chromatograph or a HPLC or you may even use simple titration methods to find the concentration of the products. And so the values can be practically continuous, usually the value beyond 3 digits after the decimal point are not reported you give only up to 2 to 3 digits depending upon the accuracy of your instrument.

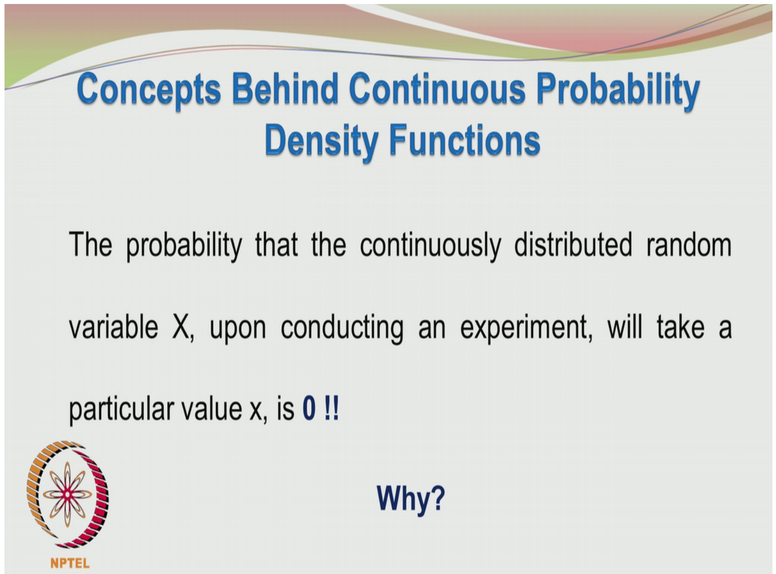
(Refer Slide Time: 03:09)



Another example of the continuous random variable is the case or situation involving the heating of water kept outside in the sun, let us take a bottle of water from the fridge it may be around 10 or 15 degree centigrade, and then we keep it out in the sun, the temperature of the water will increase and if you are measuring the temperature with thermocouple, it can show values that are anywhere between let us say 15 degree centigrade to let us say 35 or 40 degree centigrade okay.

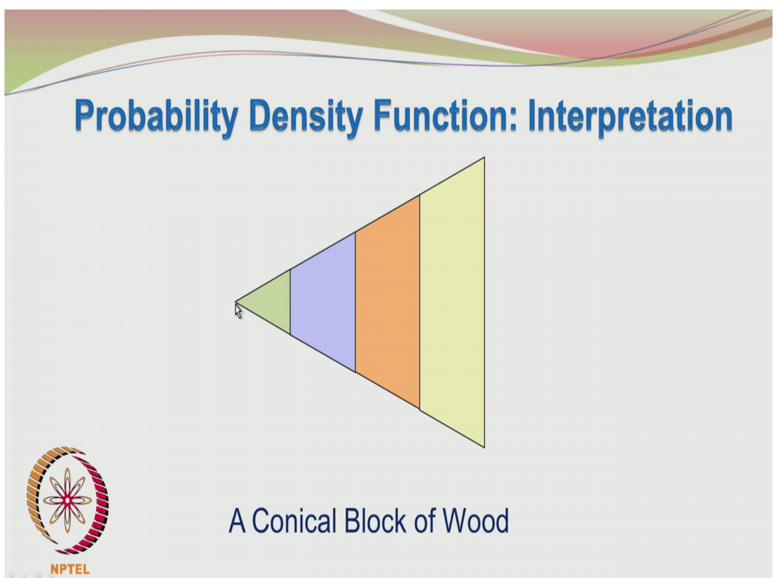
And these values are continuous, the thermocouple depending upon the accuracy may be reporting up to 2 decimal places. So the point here is we are not having discrete temperature values, but the temperature values varying over the range between 15 degree centigrade to say 35 or 40 degree centigrade.

(Refer Slide Time: 04:17)



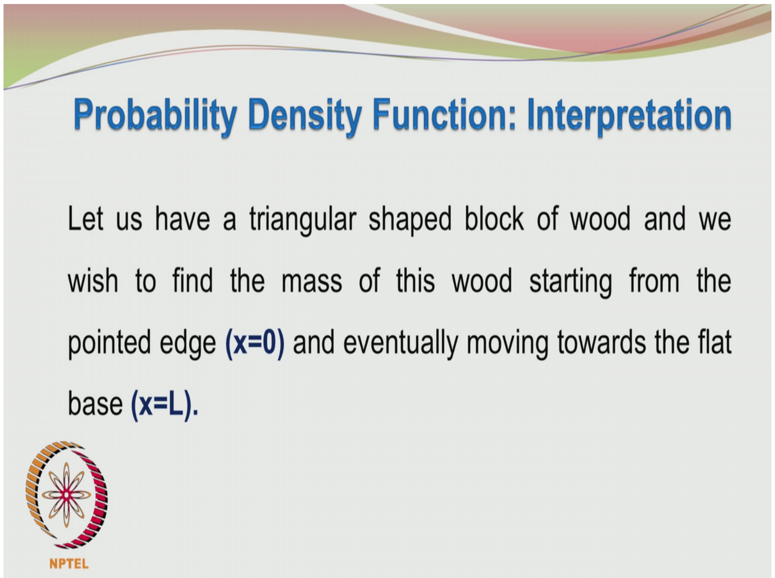
We cannot talk of the probability of a random variable taking a particular value X in the case of continuous probability distributions okay, here the probability of the random variable taking a specific value within the range is actually 0, I am going to give the reason for it.

(Refer Slide Time: 04:51)



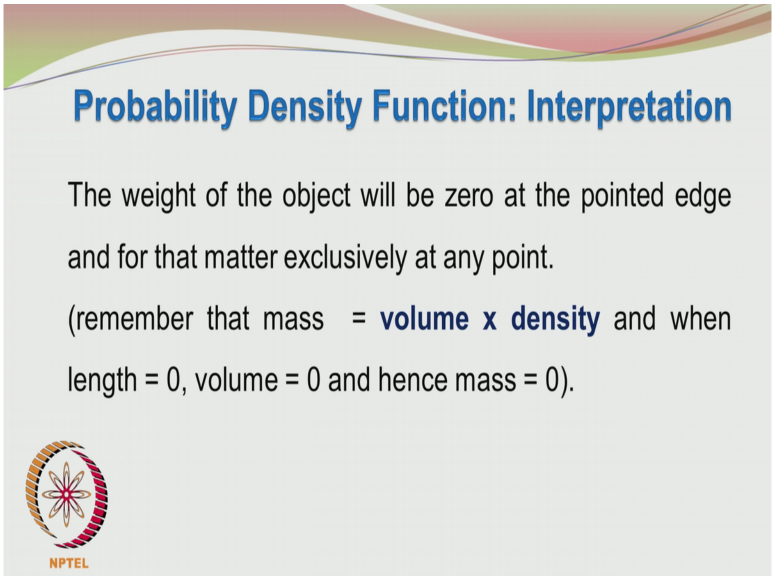
Let us consider a conical block of wood, which is having different colored sections, you can see that it is a terminating in a point, and then expanding to the base. Now if you want to find the weight of the wood at the point okay, obviously the weight of the wood at the point is 0, but if you take a certain section of the wood let us say covered by the green portion it will have a certain weight. If you add up all the colored portions, you are going to get the total weight of the wood.

(Refer Slide Time: 05:31)



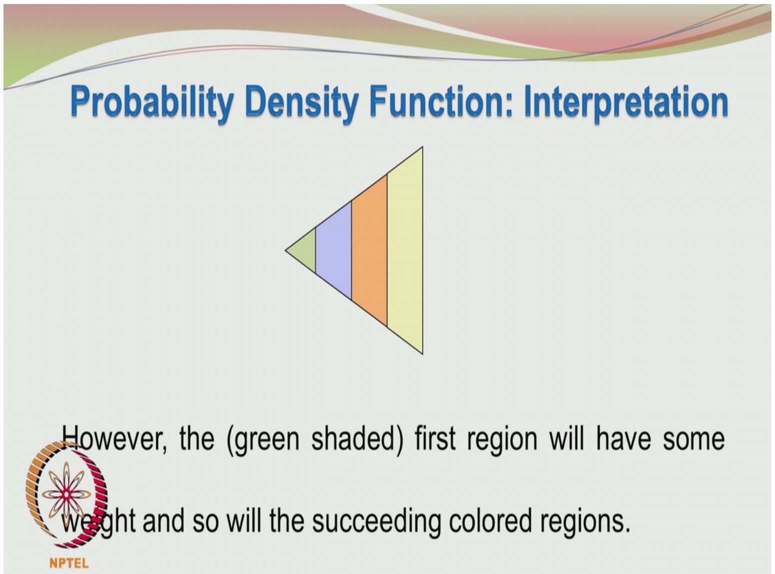
So we want to find the mass of the wood starting from the point at edge x=0, and eventually moving towards the flat base x=l.

(Refer Slide Time: 05:45)



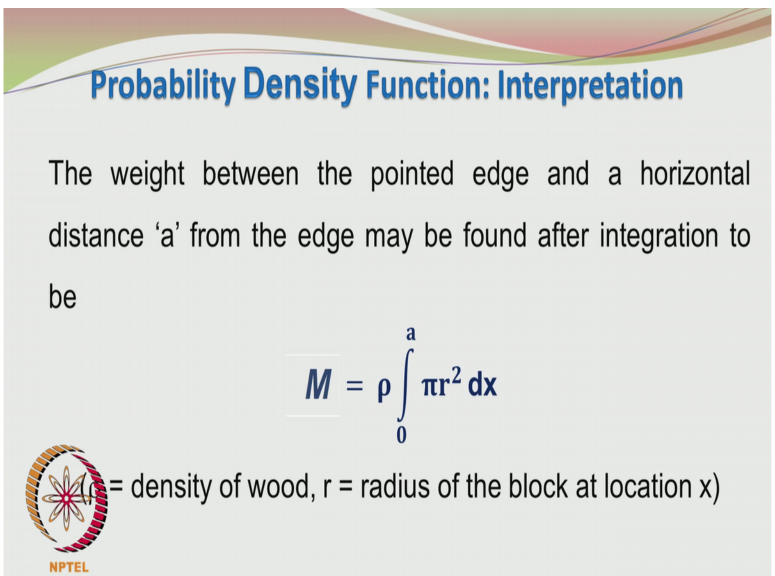
Since the mass =volume times density, when the length=0 the volume is also=0, and hence the mass will be 0.

(Refer Slide Time: 06:05)



The green portion however, occupies a certain length and it occupies a certain volume, and so it will have a certain mass.

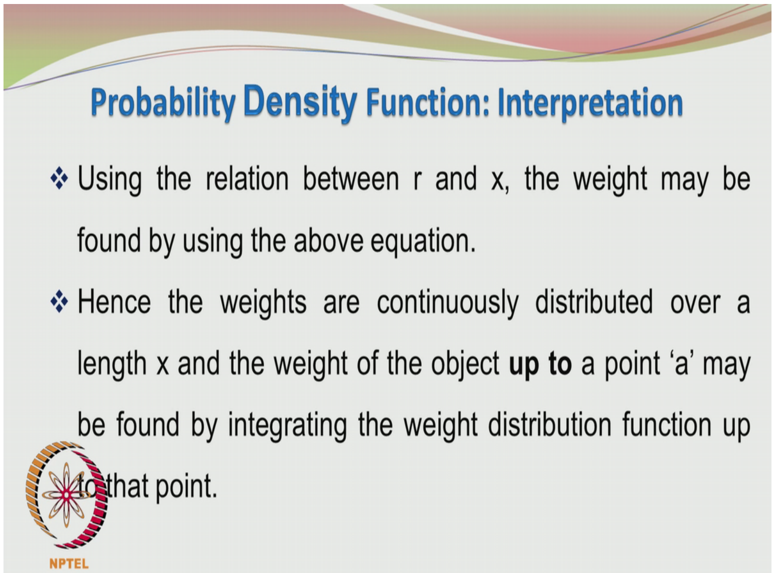
(Refer Slide Time: 06:15)



Let us take any distance from the pointed edge okay, then the distance be horizontal and let it be denoted by a, the mass of the conical block of wood until the point a is given by m =rho*integral of 0 to a pi r square dx okay. The need for integration arises, because r is changing with x, the radius of the conical block of wood is varying with distance, so you would like to do integration and find the mass, a can be anywhere between the conical tip to the base.

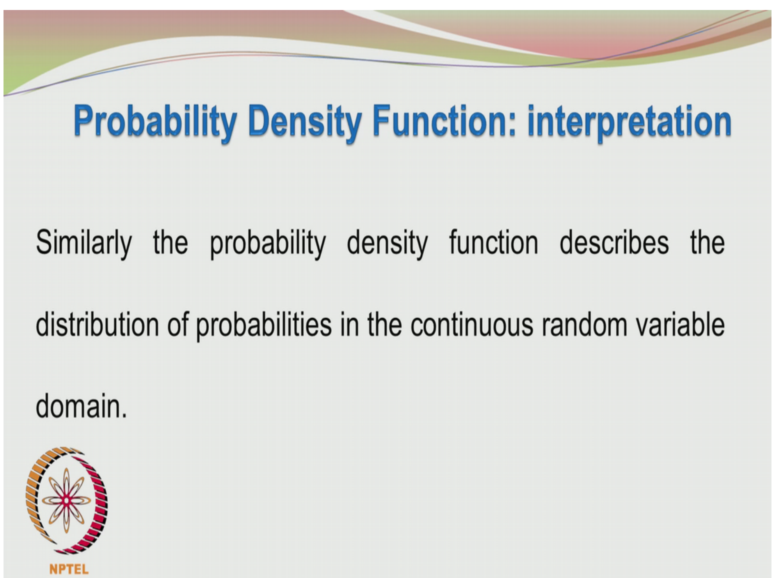
In order to carry out this integration we need to find out the relation between the radius and the x the distance from the tip okay, and that can be easily found.

(Refer Slide Time: 07:27)



The important thing to note here is the weights are continuously distributed over a length x, and the weight of the object up to a point a may be found by integrating the weight distribution function up to that point.

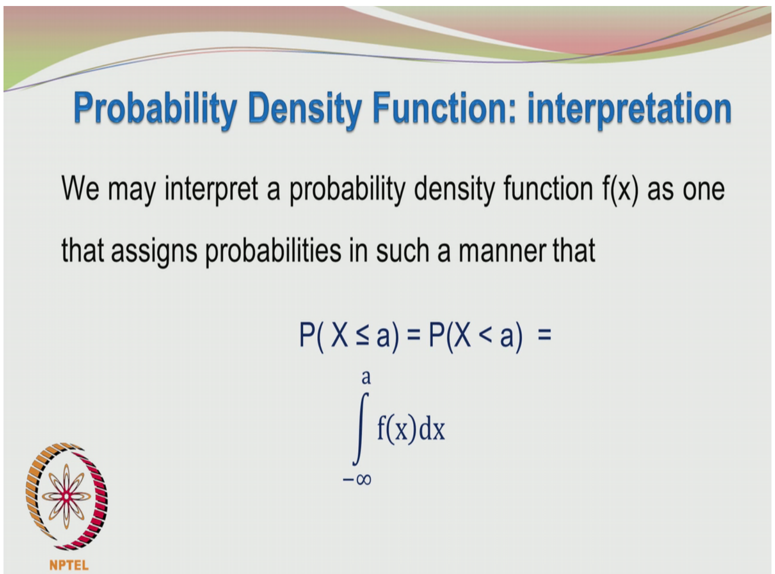
(Refer Slide Time: 07:47)



The same concept we can use to describe the probability density function; it describes the distribution of probabilities in the continuous random variable domain okay. So the probability that the random variable x will take a particular value within that range is actually 0, just as the weight of the wood at a particular point is 0. So in order to find the weight of the wood, we have to take a certain portion of the wood.

In the same way in continuous probability distributions, we want to find the probability of the random variable x falling between 2 values within the range okay, you can even start from the lower limit of the continuous random variable range up to certain point in the same range, you will have non 0 probability.

(Refer Slide Time: 09:05)

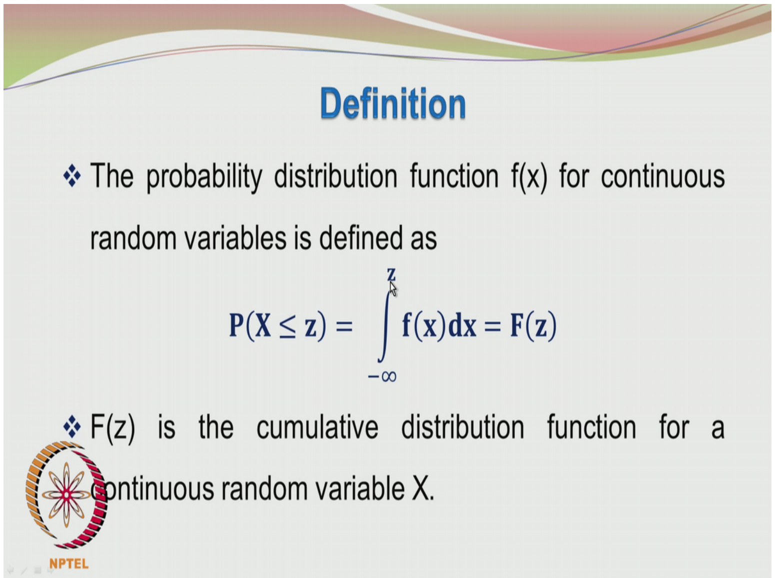


We may interpret a probability density function f of x has 1 that assigns probabilities in such a manner that probability of X<=a = probability of X <a is given by -infinity to a f of x dx. So we are doing the integration here, and if you look at the slide it can be seen that we have probability of X<=a or probability of X<a, it is immaterial whether we put X<=a or X<a. Since the probability of the random variable taking a particular value okay a specific value 0.

It is immaterial whether we use the <= sign or < sign, we have the probability density function or the probability distribution function f of x, we then integrate it between the lower limit to the required value a. In many particle situations the random variable may not take a value of -infinity for example if you are crushing the rock, the smallest particle size maybe very very small in the order of let us say 1 micrometer okay, so definitely it is not -infinity it is not even negative.

But in some cases we take the logarithm of such values and so it may become negative right, the upper limit also need not be + infinity it may be a finite value, so if you are having a finite lower limit and the finite upper limit the random variables value beyond these limits will be definitely 0 okay. It will take only values within the specified interval, it is assumed that beyond this interval the random variable value is going to be 0.

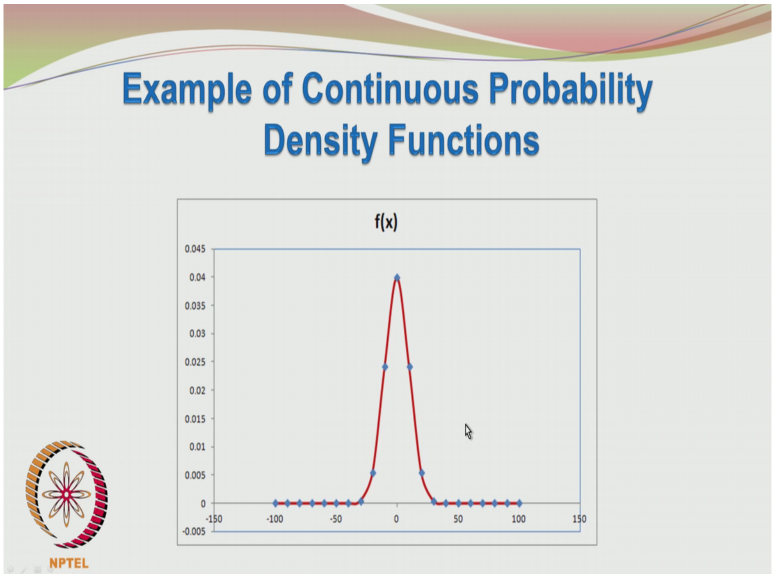
(Refer Slide Time: 11:27)



So again coming to the definition of the probability density function, probability of the random variable taking a value <= z, so we put that z in the upper limit of the integral sign, and then we call it as the f of z okay. This f of z is also a familiar quantity to us, it is nothing but the cumulative distribution function for the continuous random variable x okay. It is as of your accumulating or aggregating the probabilities up to z, and so it is called as the cumulative distribution function.

Earlier, we were using the sigma sign for adding up the probabilities, now we are using the integral sign.

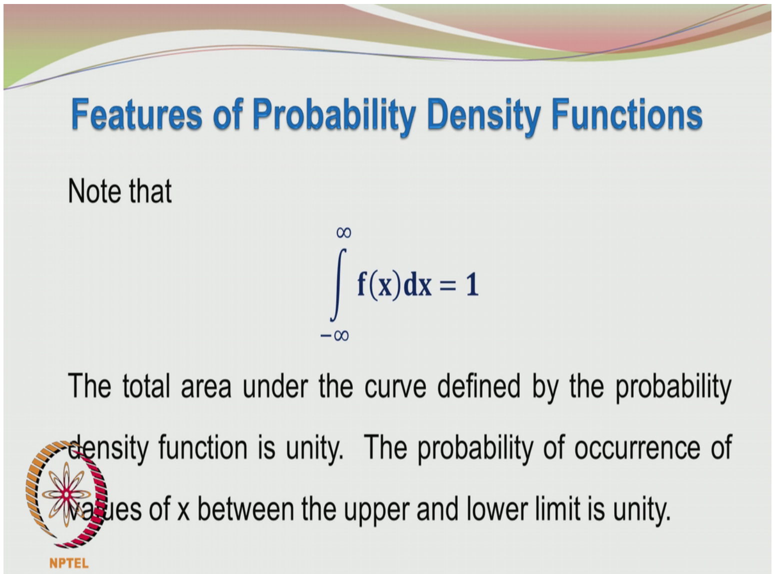
(Refer Slide Time: 12:22)



This is an example of a continuous probability density function, you can see that beyond -100 and +100 there are literally no points okay there are no representation of the data, which means that your lower limit is -100 and the upper limit is +100. Even these values up to 100 on either side of the origin are pretty small, but once you reach let us say about 35 or so, the values start to increase, and they reach a maximum value at the origin.

This probability density function is symmetry in nature, because the area under the curve on either side of the origin is the same, for example if I take 0 to -10 I find the area under the curve that will be same as the area under the curve from 0 to +10. This is an example of a symmetric distribution, but not all distributions need to be symmetric, some of the distributions may have skewness they may be having an orientation towards the lower values or a preference for lower values. And so they may be having a peak earlier, and then they may have a long tail.

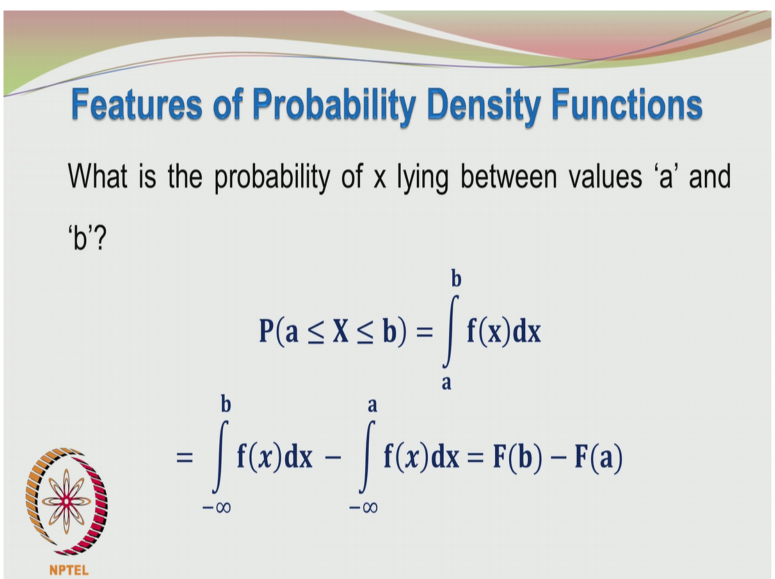
(Refer Slide Time: 14:00)



Let us look at the some of the features of the probability density functions, the area under the curve=1. If you recall the discussion for discrete probability distributions sigma f of xi was=1, which means that if you add up all the random variables probabilities, then they should add up to 1. Similarly, when you take the area under the curve which represents the probability distribution you should get 1 okay.

So essentially we are saying, what is the probability of the occurrence of the values of x between the upper and lower limit okay, we have been including all the values between the upper and lower limit and the probability will be=1.

(Refer Slide Time: 15:03)

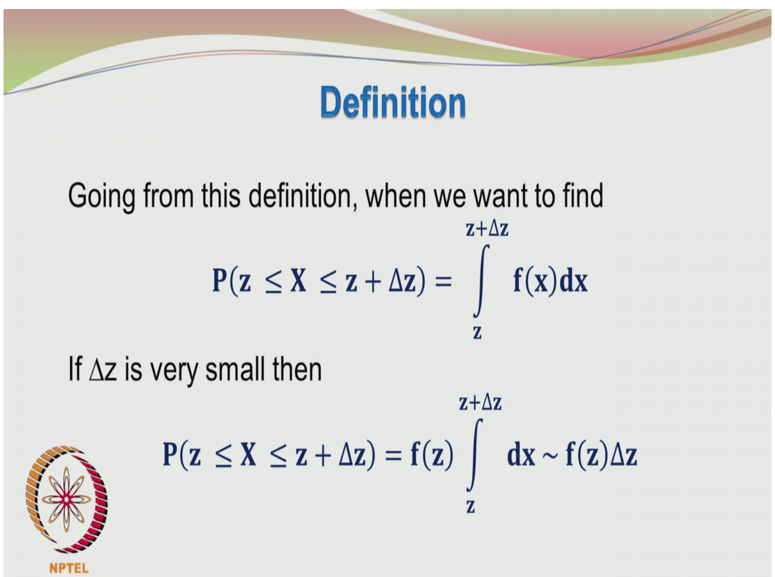


So we are really more interested in finding out the probabilities of the random variable lying between 2 values or 2 numbers, here we are more interested in a specific value a, a specific value b, which do not correspond to the lower limit or upper limit okay. We can sometimes relax the restriction we can even put a as the lower limit, or b as the upper limit, but both of them may not be the limiting values. If your range is between let us say 0 to 10 okay.

We may be interested in finding the probability of the random variable taking a value between 2 and 4 or any value okay. So in such a case the answer is !=1, it will be a value lower than 1, for that we need to do the integration we take the limit a to b for the integration, and then integrate f of X between these 2 limits. This may be written as -infinity to b f of x dx- -infinity to a f of x dx that will give us the area under the curve between the a and b.

And that is nothing but the difference between the 2 cumulative distribution functions, the cumulative distribution function evaluated at b, and the cumulative distribution function evaluated at a, so we get F of b-F of a.

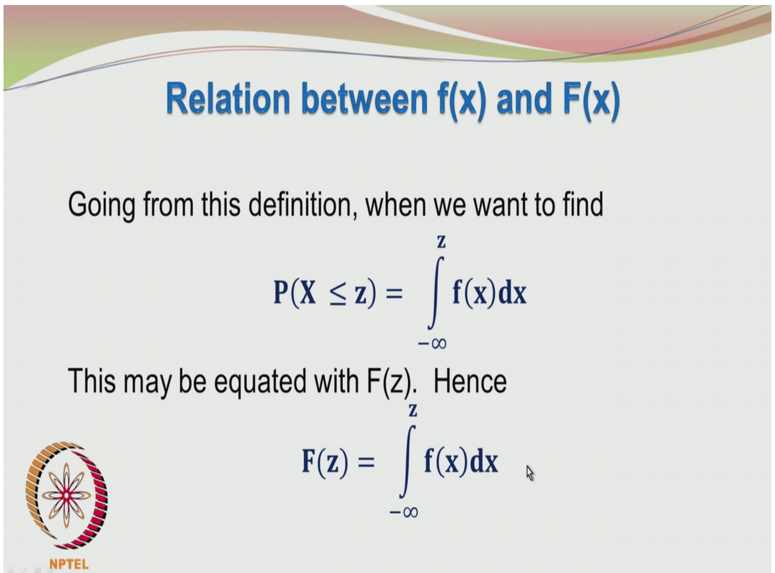
(Refer Slide Time: 16:51)



Suppose the 2 numbers a and b are such that they are very close to each other, let us say that 1 value is z and the other value is z +delta z, we want to find the probability between z and z +delta z. So what we do is we integrate between z and z +delta z f of x dx. If delta z is very small okay, then we can write probability of z <=X<=z +delta z as f of z times delta z, in the small range of z to z +delta z f of x, x is the value f of z and it does not change significantly in this small interval.

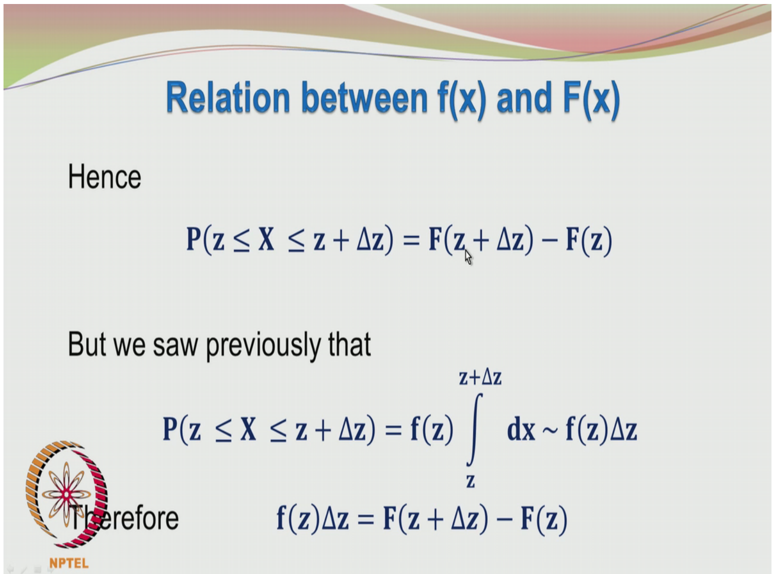
So we may as well right it as a f of z, and since it is constant between these 2 values or pretty much constant between these 2 values, we may take it outside the integral sign and then simply evaluate dx between z and z +delta z, and we get f of z delta z. Please remember this result we will be using it shortly.

(Refer Slide Time: 18:17)



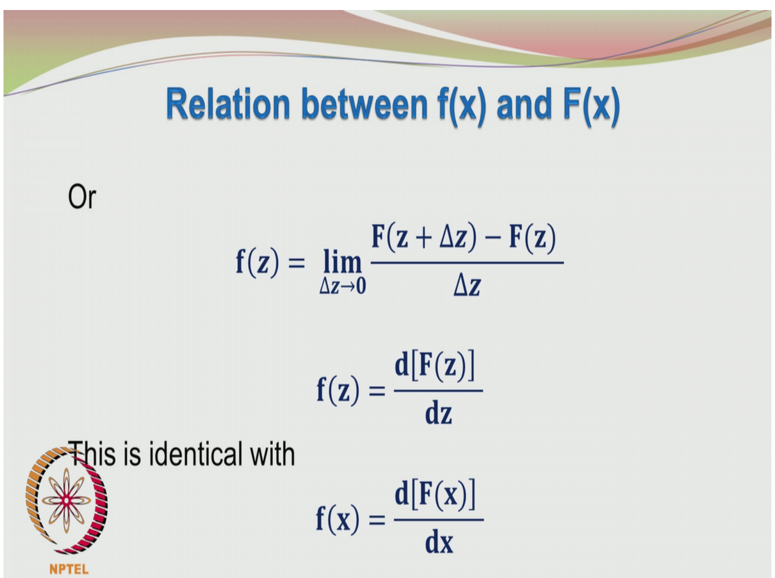
Now we want to find the relation between the probability density function and the cumulative probability distribution function, in other words we have to find a relation between small f of x and capital F of x. So the definition for the cumulative distribution function corresponding to z +delta z that is probability of X<= z +delta z= -infinity to z +delta z f of x dx. This is nothing but f of z +delta z, so we can write it as f of z +delta z -infinity to z +delta z f of x dx. We also know on the same lines that f of z= -infinity to z f of x dx.

(Refer Slide Time: 19:22)



Hence the probability of random variable X lying between z and z +delta z may be written as F of z +delta z-F of z, the last 2 F's are capital F's representing the cumulative probability distribution. However, just a brief while ago, we saw that probability of z<=X<=z+ delta z was more or less = f of z delta z. So we may write f of z delta z as the difference between the 2 cumulative probability distributions okay. So f of z delta z= F of z +delta z-F of z.

(Refer Slide Time: 20:19)

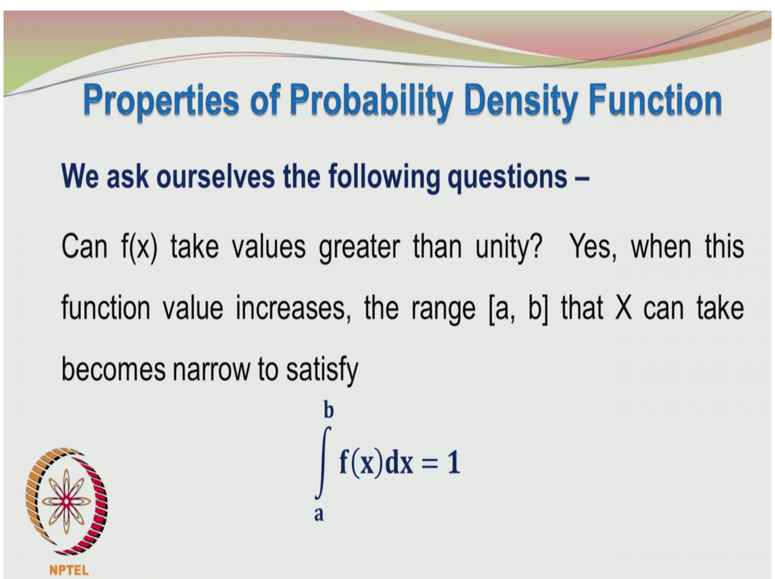


We can now write after dividing by delta z, f of z=limit of delta z tending to 0 F of z +delta z-F of z/delta z, and that= f of z= d F of z/dz okay, we are now taking the derivative of the cumulative distribution function with respect to z, then we retrieve or get the continuous probability density okay. It is quite simple in fact you please recall that when we wanted to get the cumulative distribution function, we carried out the integration okay.

Now when you want to get the probability density function from the cumulative distribution function, we have to carry out the differentiation, we differentiate the cumulative difference distribution function f of z with respect to z to get the small f of z. Change in the variable from z to x, we get f of x =d Capital f of x/dx, so when you are given the cumulative distribution function you can obtain the probability distribution function or the probability density function by simple differentiation.

When the cumulative distribution function is given in the form of a equation okay, you can differentiate this equation with respect to independent variable and retrieve the continuous probability density function, you can retrieve the continuous probability density function. Sometimes, you are going to get only experimental data, you may not have a mathematical model to that equation, then you may have to carry out numerical differentiation to get the form or shape of the probability density function, these are simple elementary calculus.

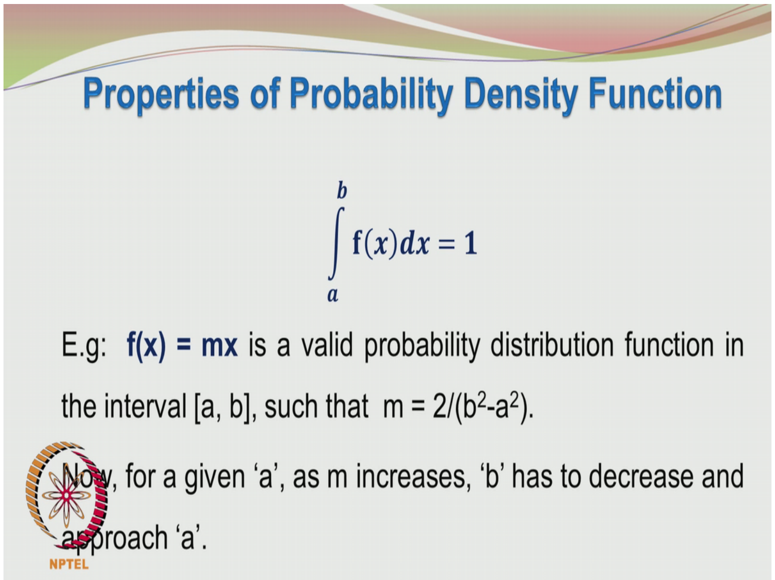
(Refer Slide Time: 22:44)



A certain query may arise when you are handling the probability density function, we know the area under the curve=1, integral of f of x dx=1, but can the f of x value itself be >unity okay. So that is a very interesting question, if the area under the curve of the function x=1 can the value of f of x exceed unity, there is no restriction on the value of f of x provided it is real okay, it should be also positive value. And beyond this there is no real restriction on the range.

You should also note that the constraint integral of a to b f of x dx=1 should be satisfied, so if f of x value takes >1 numbers, then the upper and lower limit should adjust accordingly, so that this constraint is satisfied.

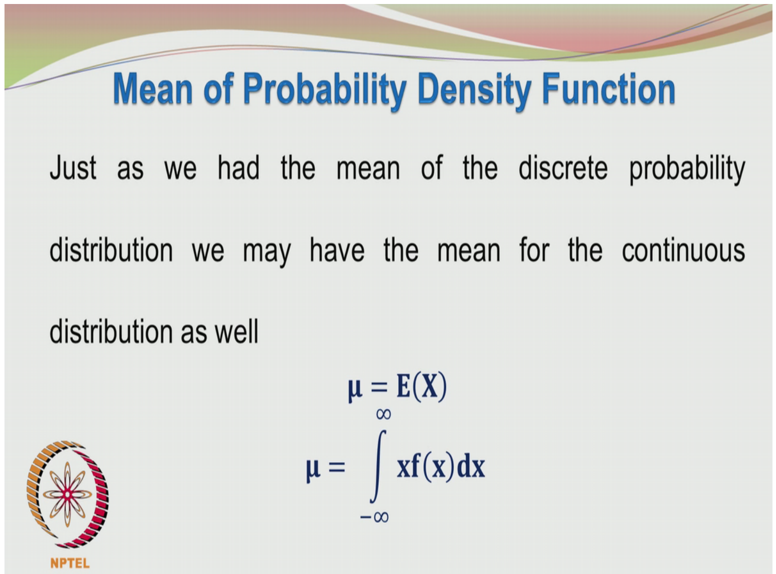
(Refer Slide Time: 24:04)



Let us take a simple case, where the function is given by mx, so when you integrate mx between the limits a to b you get mx square/2 or m/2*b square-a square, and since that will be=1, m/2*b square-a square=1, m will be=2/b square-a square, when the value of m increases okay the b will decrease and approach a. So that the constraint of the area under the curve being =unity is satisfied.

To demonstrate this with respect to this figure, if the value of f of x keeps increasing or goes to higher value, then the width will start to shrink, and the lower limit and the upper limit will start approaching each other. So when the curve tries to go upwards it sort of becomes thinner and thinner okay, the limits approach each other.

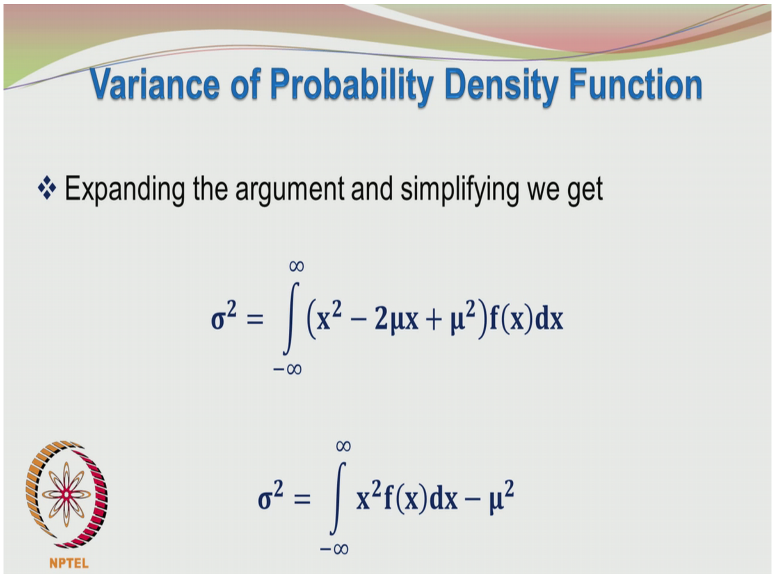
(Refer Slide Time: 25:27)



In the discrete probability distribution analysis, we came across the mean mu and the variance sigma square, how are these parameters defined for the case of the continuous random variable? In fact the definitions are quite similar, we use integration sign here while we use the summation sign for the discrete case. Let us look at the mean of the probability density function mu =the expected value of a random variable x okay, so we have mu= -infinity to +infinity x f of x dx.

This is the definition for the mean or the average value for the continuous distribution, we can also define the variance of the probability density function. The variance is defined as the expected value of x-mu whole square, here x is the random variable, for continuous distributions we have sigma squared= -infinity to +infinity x-mu whole square f of x dx.

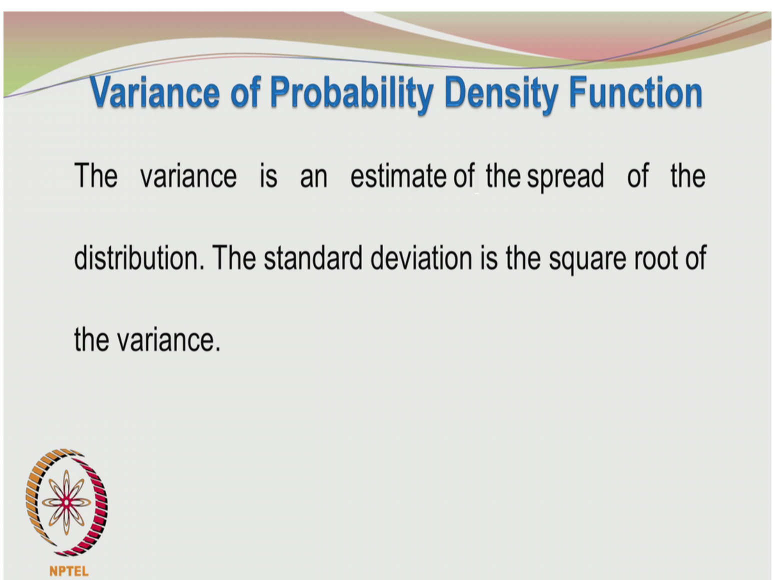
(Refer Slide Time: 26:54)



We can expand the x-mu whole square and try to simplify this integral, so we get x squared-2 mu x +mu square f of x dx, and after a bit of simple manipulations, we get sigma square= -infinity to +infinity x square f of x dx-mu square. What happened here is we multiply f of x with x square and take it out we have a separate integral -infinity to +infinity x square f of x dx -2 mu x f of x dx became -2 mu square using the definition for the mean.

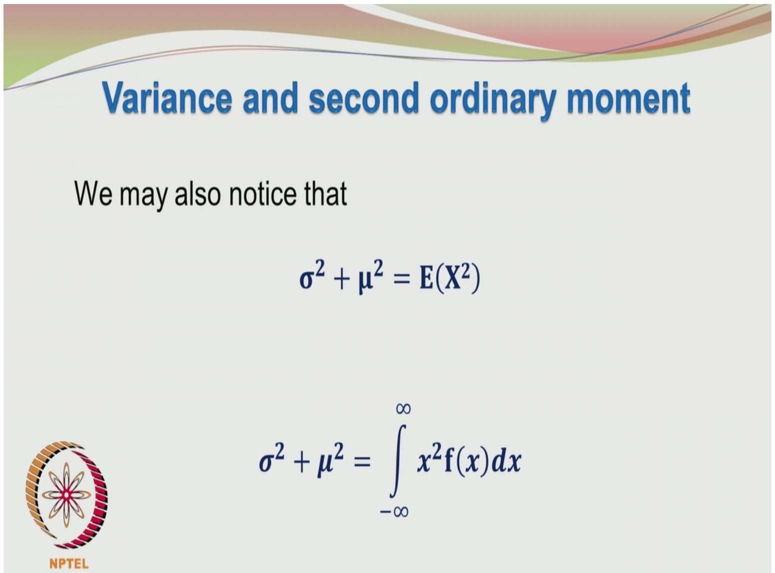
And then you have mu square f of x dx, when you integrate that -infinity to +infinity f of x dx =1, so you get -2 mu square+ mu square which reduces to -mu square, so sigma square the variance of the probability density function= -infinity to +infinity x square f of x dx-mu square.

(Refer Slide Time: 28:21)



The variance is an estimate of the spread of the distribution, to find the standard deviation use take the square root of the variance and report the positive value.

(Refer Slide Time: 28:36)



By now you may have also noticed that sigma square +mu square is expected value of x squared, this result we have also seen in the case of discrete probability distributions, E of x square by definition is -infinity to +infinity x square f of x dx. These are parameters, which we will be frequently coming across in continuous probability distributions, some of you may be having the apprehension perhaps that every time you have may have to do the integration, to find the mean and variance.

Usually, we will be working with standard probability distributions for which the mean and variance are well known, so we do not have really have to do the integration every time. The next query is suppose I want to find the probability of the random variable x lying between 2 numbers okay, should I do the integration? For standard probability distributions tables of data are available for finding out these probabilities.

In case the function is unknown or a new function which has not been studied previously, then of course you will have to carry out integration, if the integration is not possible analytically you may have to carry out the integration numerically. Nowadays, the software's are so advanced that either carrying out the numerical integration or even the analytical integration is not a big task.

If you look at standard software spreadsheets, numerical computation programs, mat lab, they all can directly give you the value of the probability when you specify the limits of the probability distribution function, the probability values are automatically calculated and given. I would encourage you to look at some of these spreadsheets and mathematical software to find out the probabilities for some standard distributions.

I would like to encourage you to look at these spreadsheets or mathematical software, you can find out using the help command, how to find the probabilities for standard distributions they are very easily calculated through simple commands. The standard probability distributions which will be studying or the normal distribution or the Gaussian distribution, the student's t distribution, the fisher distribution and the Chi-square distribution.

There are also other distributions like the beta distribution, the gamma distribution, the viable distribution and so on, but once you have had the background for these standard distributions which I am going to talk about, you can very easily understand the other ones. So right now we will conclude our discussion on the continuous probability distributions, and go on to the most important distribution the normal distribution.