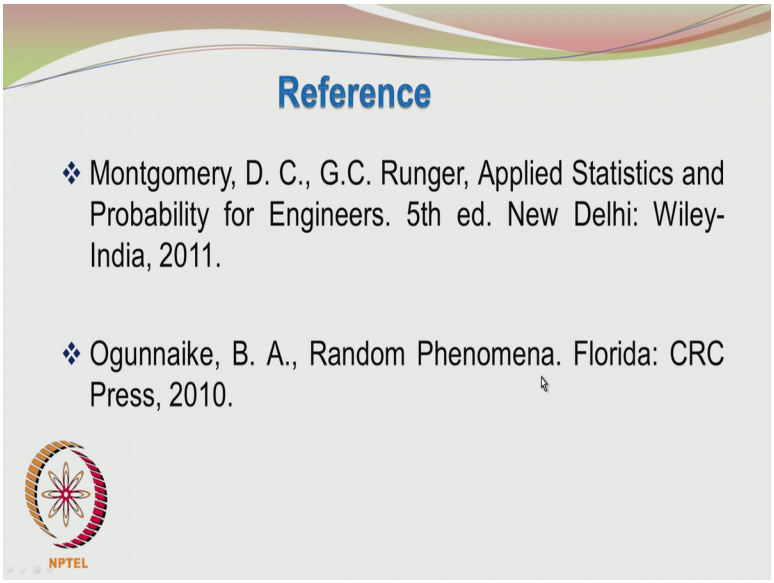
Statistics for Experimentalists
Prof. Kannan. A
Department of Chemical Engineering
Indian Institute of Technology - Madras

Lecture - 06
Normal Probability Distributions

Okay, in today's lecture we will be looking at the normal distribution, it is also called as the Gaussian distribution.

(Refer Slide Time: 00:26)



The reference books for this topic are given in this slide, the book by Montgomery and Runger Applied Statistics and Probability for Engineers, and the book Random Phenomena by Ogunnaike are suitable references for the material we are going to cover in today's lecture. So why should we study normal probability distribution, it is very popular, very elegant and relatively simple, it has some nice properties to it.

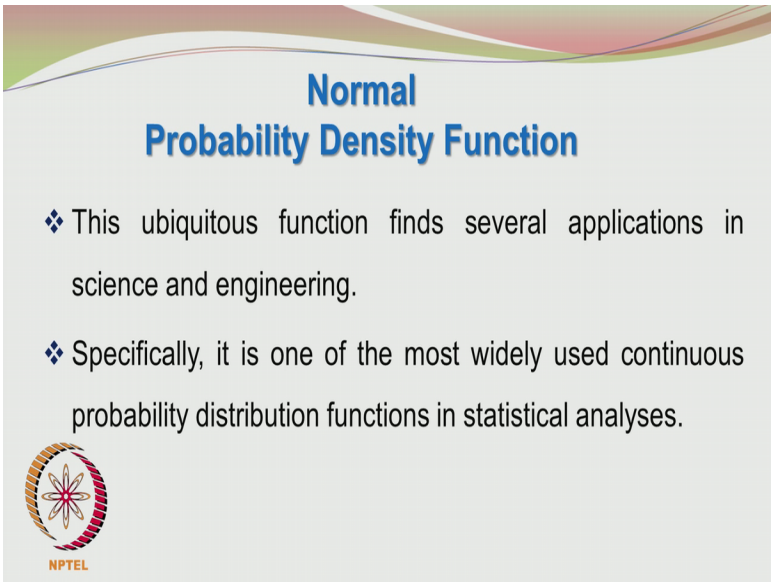
We use normal probability distribution not only because of these desirable features, but it is also the distribution which real life tends to follow in many cases, in large classes the distribution of marks is considered or approximated to be a Gaussian or normal distribution. If you look at the particle sizes coming from a crusher or a grinder they may cover a very large range of values, the smallest particle maybe in the micron range, and the largest particle maybe in the millimeter or centimeter range.

These sizes let us denote them by D, you take the natural log of the particle sizes convert them into ln D, you will be surprised to find that the distribution of the natural logarithm of the particle diameters is following the normal probability distribution. Once you have this normal probability distribution, you can do a lot of things, in the case of the marks distribution you can find the percentage of the students, who have got marks let us say between 50 to 60.

You can also find out what is the percentage of students, who have got marks below 20, regarding the particle sizes you can also calculate the probabilities that the particle sizes are going to lie between 2 values. It is a very useful distribution, it is also going to be useful for our design of experiments and analysis of data, because many of the samples they have properties like mean and variance and so on.

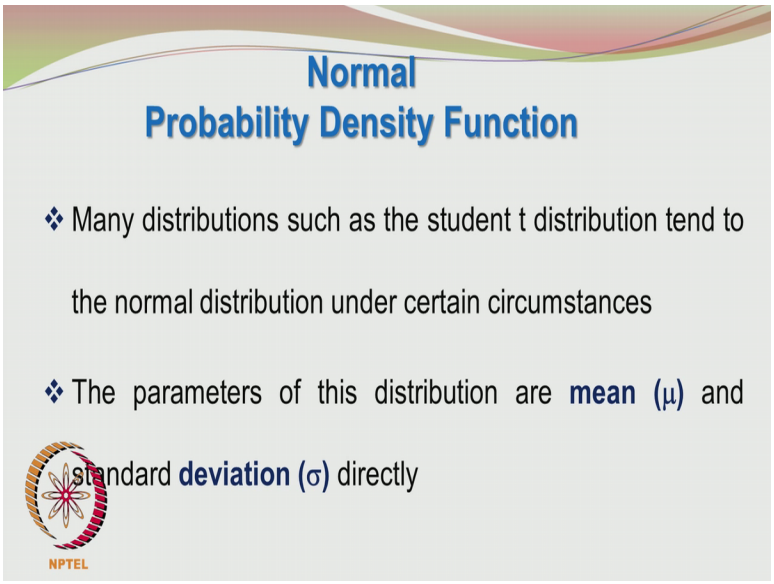
If you look at the mean values of different samples, sometimes they follow the normal probability distribution okay. Some of the other standard distributions like the t distributions also tend to normal distribution under certain conditions.

(Refer Slide Time: 04:16)



This probability density function finds several applications in Science and Engineering, it is one of the most widely used continuous probability distribution functions in statistical analysis.

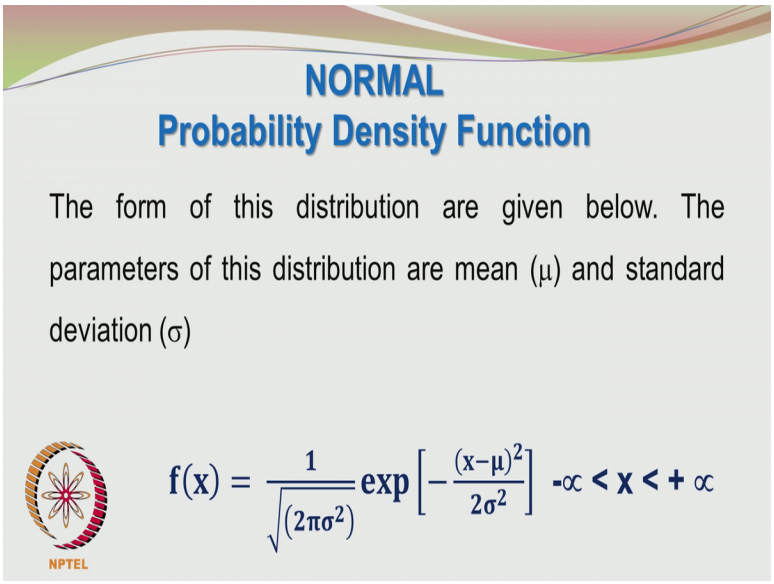
(Refer Slide Time: 04:36)



When we are looking at the parameters of the normal probability distribution, it is very interesting to note that the parameters of the distribution are themselves the mean and standard deviation of the distribution, this is a big advantage. In many other distributions you will have 2 parameters or in more infrequent situations you may have even 3 parameters. But in the case of a normal distribution you have 2 parameters.

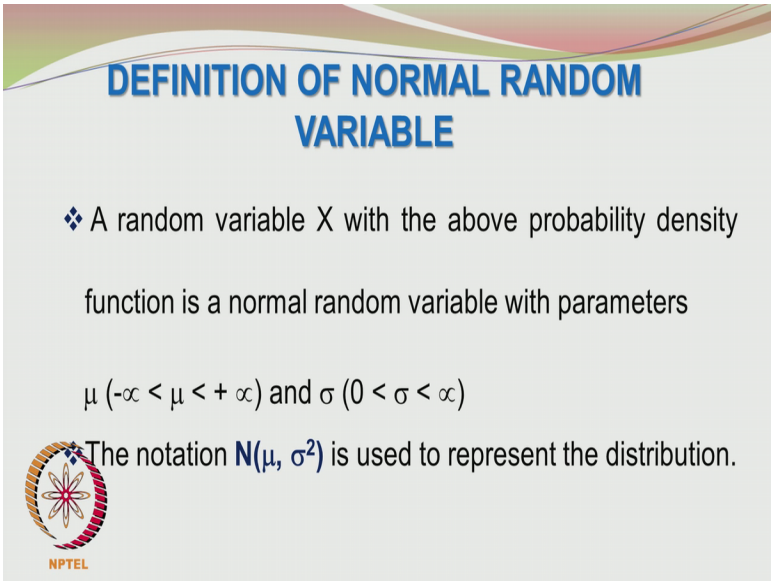
The 2 parameters describe the shape of the distribution, and the parameters are themselves the mean and standard deviation, for other distributions you have some parameters let us say we call them as parameter 1 and parameter 2, these 2 parameters are then used in a mathematical expression to get the mean and variance.

(Refer Slide Time: 06:00)



Let us look at the probability density function for the normal distribution, the function is denoted by f of x and that is given by 1/root 2 pi sigma square exponential-x-mu whole square/2 sigma square, mu is the mean of the distributions, sigma squared is the variance of the distribution, the lower limit is -infinity, and upper limit is +infinity, so x can take both negative values as well as positive values.

(Refer Slide Time: 06:44)



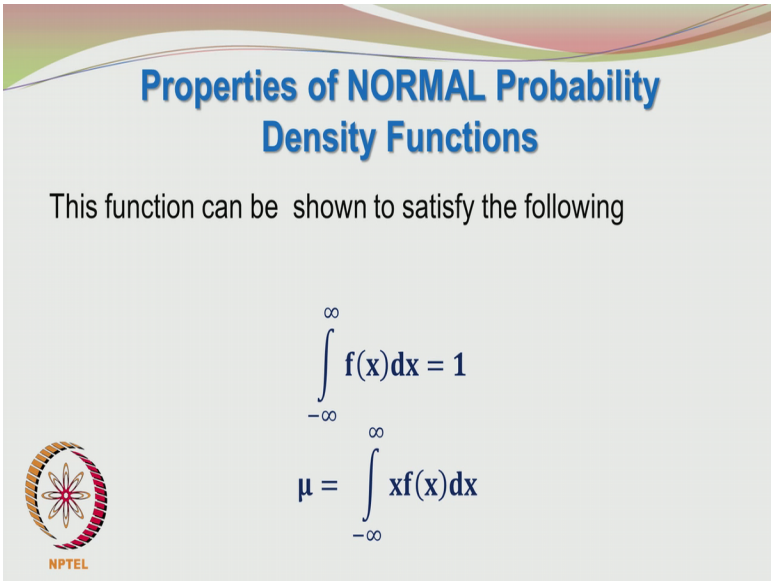
Now we will define the normal random variable, a random variable X with the above probability density function is a normal random variable with parameters mu and sigma, the parameter values may range from -infinity to +infinity for the case of the mean, and the case of the standard

deviation it may vary between 0 to infinity. This is an important point many of us intuitively believe that mean value of a distribution should be only positive.

It need not be the case mean is an average okay, and so the distribution range may be such that the mean value or the average value may be negative. For example, if the upper and lower limits of the distribution are -5 to -50, then the mean would be somewhere between -5 and-50, so the mean can take negative values. However, the standard deviation is obtained as the positive square root of the variance, and so it is always going to be a positive quantity.

Coming back to the normal distribution, for a distribution with parameters mu and sigma square we use a general notation N of mu sigma square to denote the normal distribution.

(Refer Slide Time: 08:38)

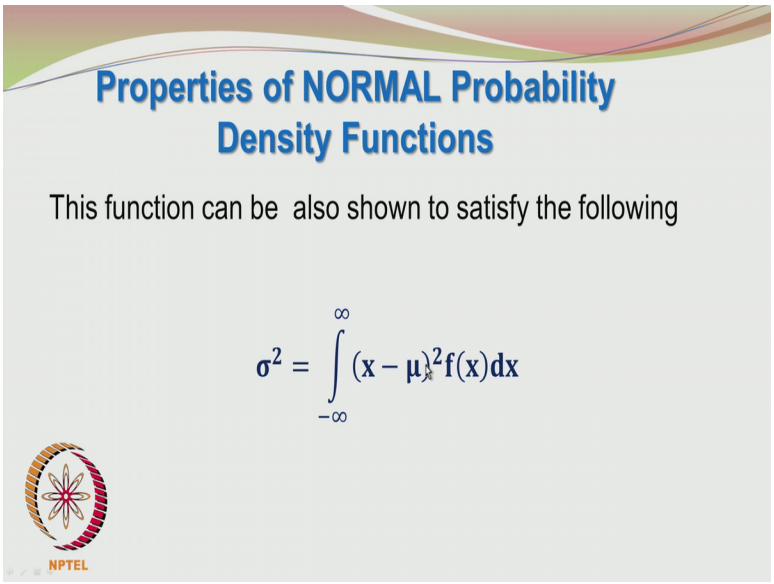


As any other probability density function, the normal distribution should satisfy the condition -infinity to +infinity f of x dx=1. In other words, if you take this expression for f of x into the integral and carry out the integration, fortunately the integration can be carried out analytically you will find that the value=1 after the application of the limits. This is not only definition for the mean mu.

But you can also show that after substituting the value for f of x the equation I just showed that this equation if you plug it here and then multiply by x carry out the integration, you will be

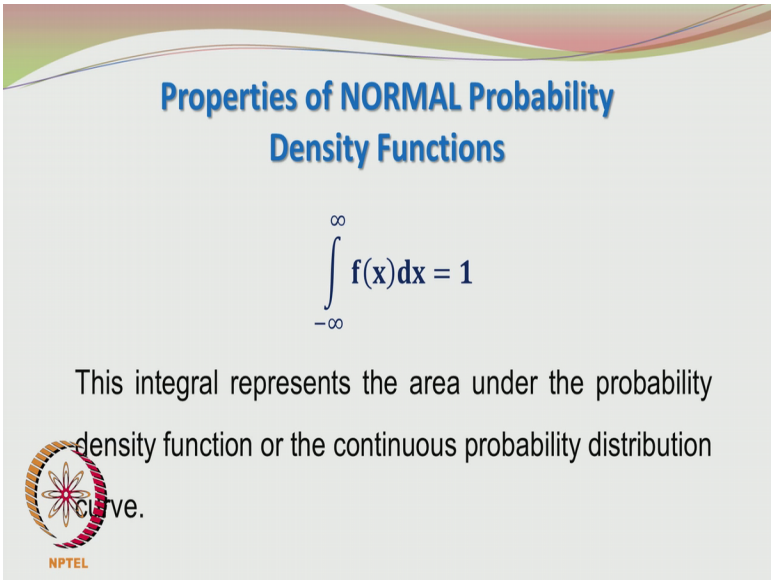
pleasantly surprised to find that value to be mu okay. The function had parameters mu and sigma square and you finally end up with mu, sigma square surprisingly will vanish in the mathematical manipulations involved.

(Refer Slide Time: 10:16)



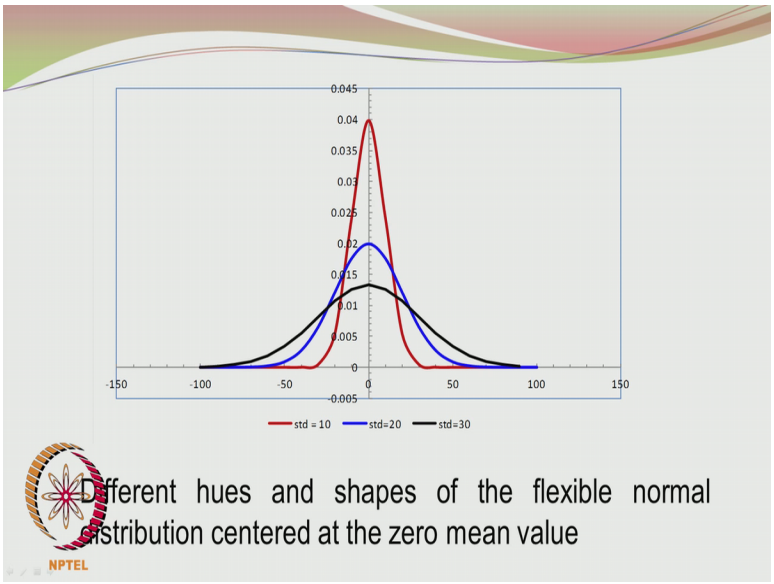
When you carry out the different type of integration okay between -infinity to +infinity x-mu whole square f of x dx, you will find after plugging in the expression for the normal distribution here, you carry out the necessary mathematical steps including integration by parts, you will find that the mu will somehow vanish, and you are going to left with only sigma square.

(Refer Slide Time: 10:50)



The area under the curve=1, the total area under the curve=1, when you integrate f of x between the lower limit to the upper limit. What you do is you plug in the equation of the normal distribution here for f of x, after carrying out the integration you will get the area to be 1.

(Refer Slide Time: 11:16)



The normal distribution is a very flexible one, it can change shape very easily when you change the parameters. If you look at the normal distribution, it is centered at the mean value, so by changing the mean value you can make it move around, instead of 0 if you say the mean to be 50, the normal distribution will shift and it will come to it will be centered around okay, modification here.

The normal distribution is a very flexible distribution, it can change shape and location pretty easily, the parameters of the distribution are mu and sigma. The distribution is centered around the mean mu; it is a symmetric distribution. When you change the value of mu from let us say 0 to 50, the distribution moves to the right, and it gets centered at 50. The spread of the distribution is governed by the parameters sigma or the standard deviation.

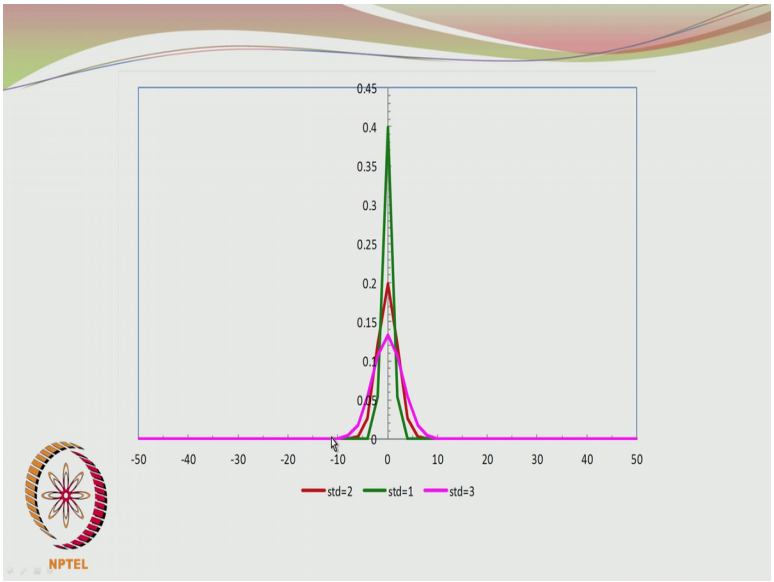
If you look at these 3 curves, all of them are normal distributions, but they have different standard deviations, they have the same mean mu of 0, but they have different standard deviations. The one showing a taller peak has a smaller standard deviation of 10, the second peak

of intermediate height has an intermediate standard deviation of 20, and the shortest peak but the widest one as well has the standard deviation of 30.

So the standard deviation is a measure of the spread, and so higher the standard deviation the more would be the spread of the distribution. You can make this normal distribution move about by changing the value of mu, so instead of 0 you can give 10 or 50 and it can just shift to this side. You can also try to imagine what would happen if you continuously decreased the value of sigma, if the value of sigma is reduced, the height of the distribution will increase.

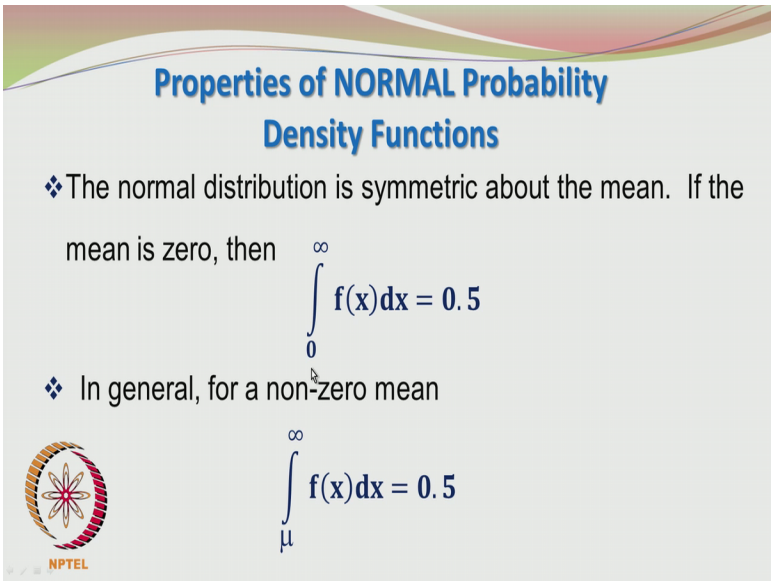
What will happen when you reduce the sigma value to 0, just think about it, what is that mathematical function called?

(Refer Slide Time: 14:23)



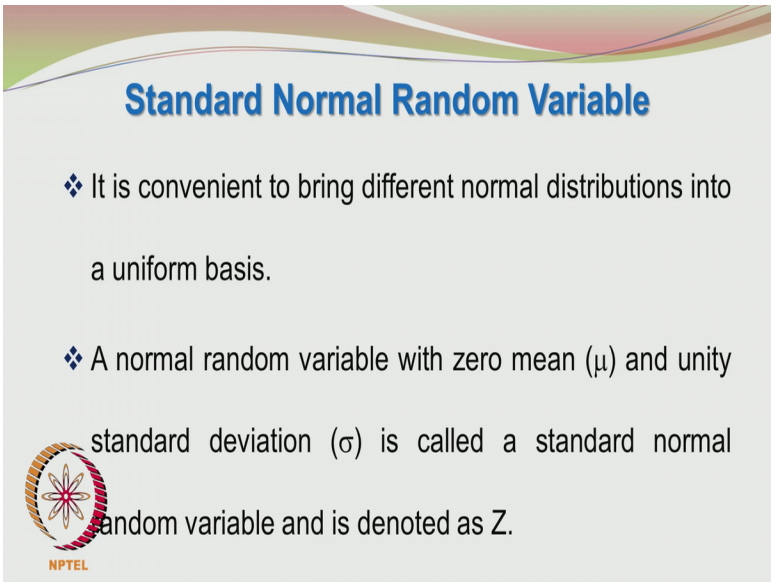
Here, we have cases where we reduce the standard deviations to smaller and smaller values, here you start with 3, the pink curve then you reduce it to 2, and then you reduce it even further to 1, the green curve shows the tallest peak. And if you keep reducing the value of sigma the f of x value will keep increasing, f of x is a maximum at the center okay. So this peak value will keep increasing, as the standard deviation decreases.

(Refer Slide Time: 15:12)



The normal distribution is symmetric about the mean, if the mean is 0 and then you cover from the mean value to the upper limit 0 to infinity, you get f of x dx =0.5. Here, we have the mean of 0, when you cover the curve from 0 to infinity along this direction, the area under the curve will be 0.5, we know that the total area under the curve=1, so when you consider half the domain from 0 to +infinity, we should get an area of 0.5. In case you have a non 0 mean, then when you integrate between mu to infinity, we get f of x dx mu infinity=0.5.

(Refer Slide Time: 16:15)



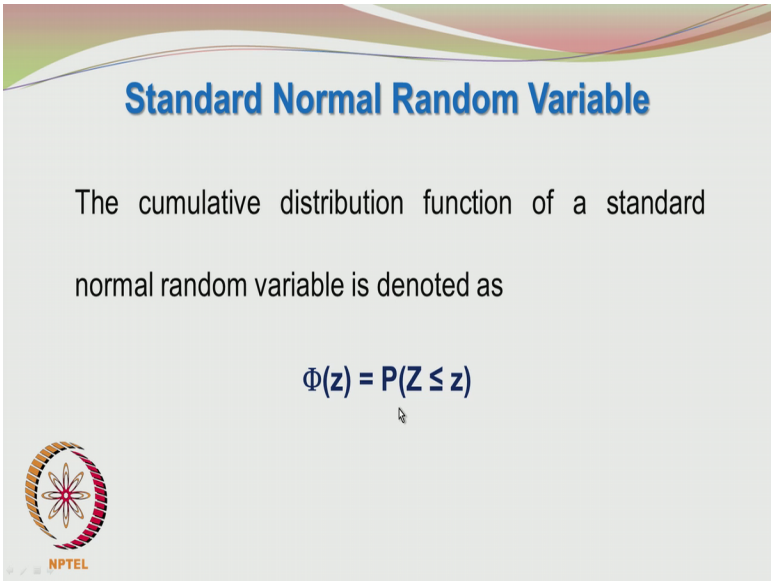
So you can have different normal distributions each one having a different mean and or variance, I told you that we do not have to do numerical integration or any further calculations to find the different probabilities involving the normal distribution. However, you can have different normal

distributions each one having different mean and or standard deviation, so for each normal distribution we cannot have a chart or table of probabilities.

It is important to reduce a given normal distribution into a standard form, the transformation of any given normal distribution to its standard form is pretty easy. Once it has been reduced to a standard form, then you need only one set of tables or charts to read the probability values. How do you define the standard form of the normal distribution? The normal distribution in its standard form has a mean of 0 and variance of 1, since variance is 1 standard deviation is also 1 okay.

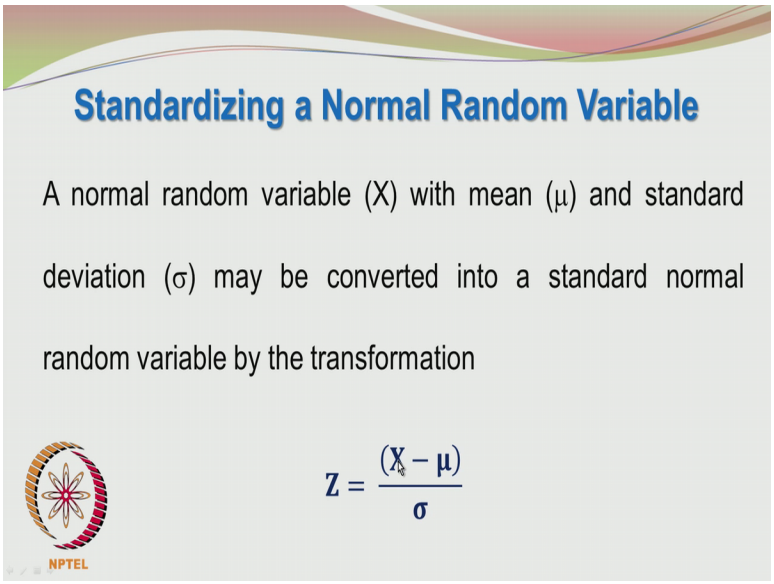
So a standard normal distribution has mean 0 and variance 1, the random variable associated with the standard normal distribution is called as the standard normal random variable.

(Refer Slide Time: 18:20)



The cumulative distribution function of a standard normal random variable is denoted as phi of z =probability of z <=z okay, we are now talking about a standard normal variable, and what is the probability that this standard normal variable will take a value smaller than that of z, and that is given by the cumulative distribution for a standard normal variable phi of z.

(Refer Slide Time: 19:00)

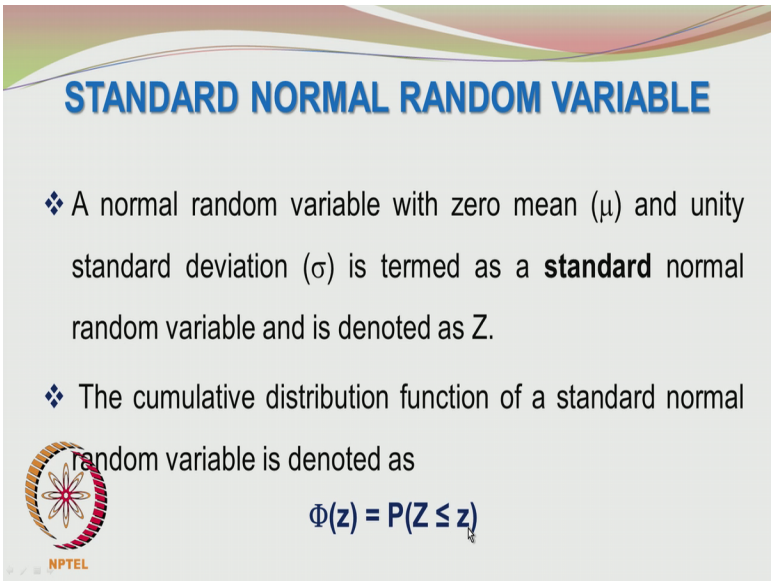


How do we make the transformation? You are having the original random variable X, which is following the normal distribution, it is having a mean mu and standard deviation sigma, you have to convert it into the standard normal random variable form. For that we do z=x-mu/sigma, x is the original random variable, mu is the original mean of the normal distribution, and sigma was the original standard deviation of the normal distribution.

By making this transformation we subtract mu from x, and then the resulting quantity is divided by sigma, we get z. After this transformation we have a new random variable z, which has a mean of 0 and standard deviation of 1, and it is also a normal distribution okay. So this transformation is applicable irrespective of the value of mu and sigma, obviously you cannot have sigma to be 0, it should be a positive number >0.

Whether it is negative mu or positive mu, it is immaterial all you have to do is carry out the transformation given by z=x-mu/sigma.

(Refer Slide Time: 20:47)

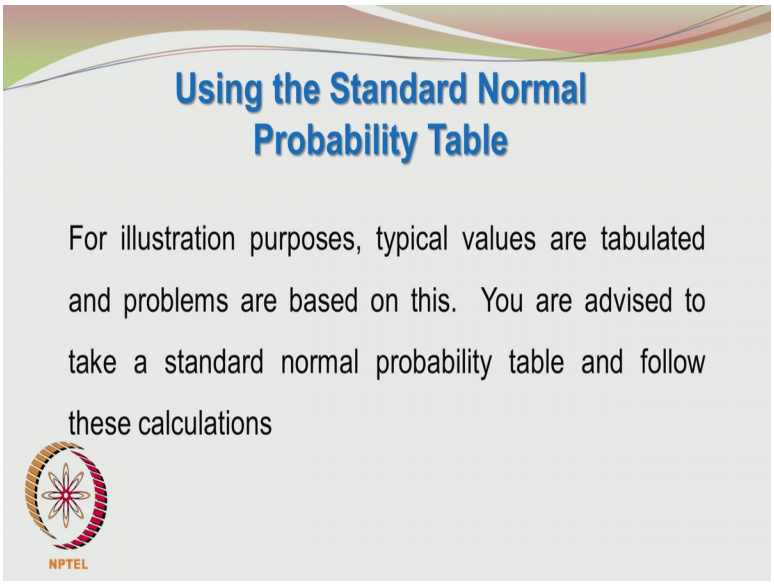


What is it really mean? When you want to find the probability of Z <=z, where z is a number the small z is a number taken by the random variable capital Z, but we do not know the value of small z initially, we only know the value taken by the random variable capital X okay, to find the small z you substitute the value of small x here, then you subtract the mean of that normal distribution, that resulting quantity you divide by the standard deviation of the normal distribution then you will get small z.

You were having the normal distribution associated with X random variable X, it was having mu and sigma as its parameters. Now when the random variable X takes on a particular value small x, you subtract the actual mean of the normal distribution from that small x divided by sigma, then you will get a value small z, this small z is used here, then you use these standard normal random variable form okay, so it becomes probability of Z<=small z.

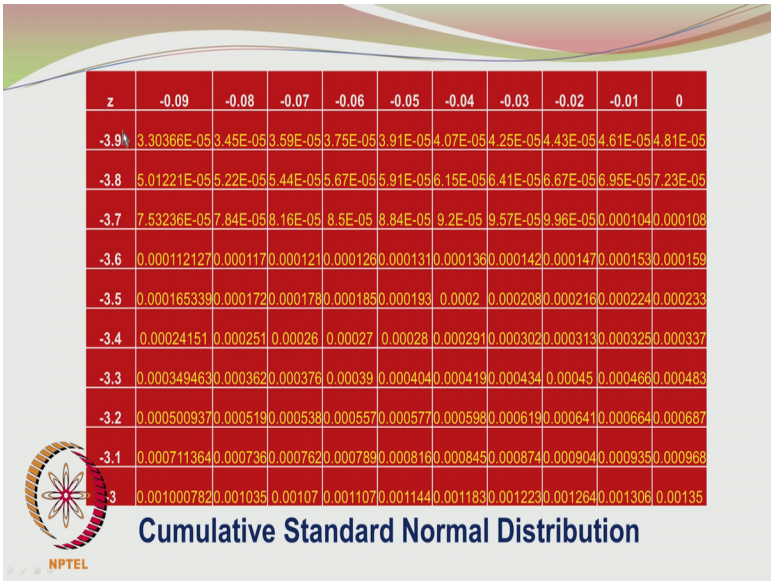
Now you can use the normal distribution with mean mu of 0 and standard deviation sigma of 1, charts are available for this particular normal distribution that is the normal distribution with 0 mean and unity standard deviation, then you can compute the probabilities.

(Refer Slide Time: 22:56)



Standard normal probability tables are available in many places including the reference book I told you at the beginning of the lecture, you can also find these tables in the internet sources, surprisingly or interestingly you can generate these tables by yourself if you have access to any standard spreadsheets. You can define values of z, and use the approximate command in the spreadsheet, and generate the complete table, I have done that.

(Refer Slide Time: 23:45)



z	-0.09	-0.08	-0.07	-0.06	-0.05	-0.04	-0.03	-0.02	-0.01	0
-3.9	3.30366E-05	3.45E-05	3.59E-05	3.75E-05	3.91E-05	4.07E-05	4.25E-05	4.43E-05	4.61E-05	4.81E-05
-3.8	5.01221E-05	5.22E-05	5.44E-05	5.67E-05	5.91E-05	6.15E-05	6.41E-05	6.67E-05	6.95E-05	7.23E-05
-3.7	7.53236E-05	7.84E-05	8.16E-05	8.5E-05	8.84E-05	9.2E-05	9.57E-05	9.96E-05	0.000104	0.000108
-3.6	0.000112127	0.000117	0.000121	0.000126	0.000131	0.000136	0.000142	0.000147	0.000153	0.000159
-3.5	0.000165339	0.000172	0.000178	0.000185	0.000193	0.0002	0.000208	0.000216	0.000224	0.000233
-3.4	0.00024151	0.000251	0.00026	0.00027	0.00028	0.000291	0.000302	0.000313	0.000325	0.000337
-3.3	0.000349463	0.000362	0.000376	0.00039	0.000404	0.00041	0.000434	0.00045	0.000466	0.000483
-3.2	0.0005009371	0.000519	0.000538	0.000557	0.000577	0.000598	0.000619	0.000641	0.000664	0.000687
-3.1	0.0007113641	0.000736	0.000762	0.000789	0.000816	0.000845	0.000874	0.000904	0.000935	0.000968
-3	0.0010007821	0.001035	0.00107	0.001107	0.001144	0.001183	0.001223	0.001264	0.001306	0.00135

So you can see this table the z value is starting from -3.9, and it is going in the vertical direction towards increasing z values. Suppose I want to find the probability corresponding to -3.75, so I locate -3.7 here, then I go originally to my right and hit the value corresponding to -3.75, so if I

go in the horizontal direction I get z values of -3.99, -3.98, -3.97 so on to -3.90. If I take a z value here and I move in the horizontal direction, I get -3.19, -3.18, -3.17 and so on.

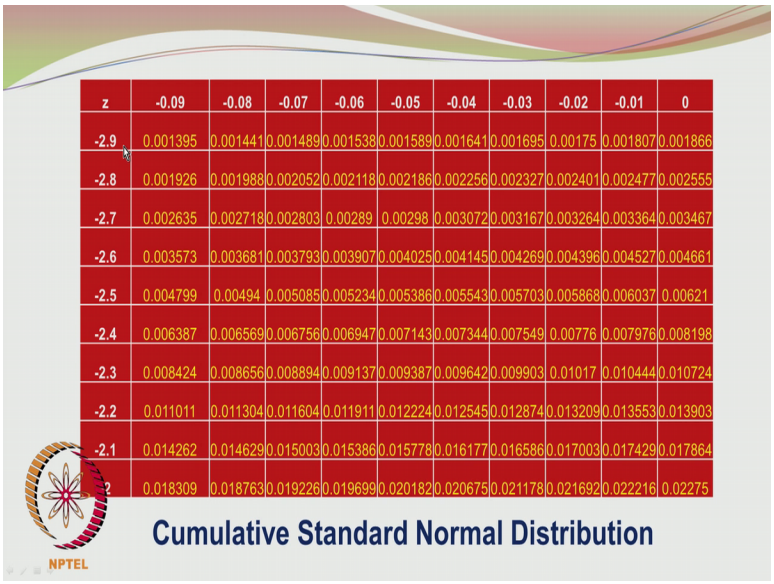
So I can read the corresponding probability values. Why we are not having values below -3.9 is the probability values or the area under the curve corresponding to-3.99 is pretty small 10 power-5, 3.3*10 power-5 which is pretty much close to 0. So if you go for z values even lower than this you will get even smaller numbers, so when you are pretty much reporting 0, it is not really necessary to report 10 power-6, 10 power-7 and so on.

Even though our range is from -infinity to +infinity, we see that the curve pretty much coincides with the x-axis at value of -4, since the distribution is symmetric the probability values or the area under the curve beyond z=+4 will also be very, very small right. So you can take any z value up to 2 digits beyond the decimal point -3.56 you just take -3.5 or locate -3.5 here, go towards your right hand side, and you will hit -3.56, -3.55 and so on.

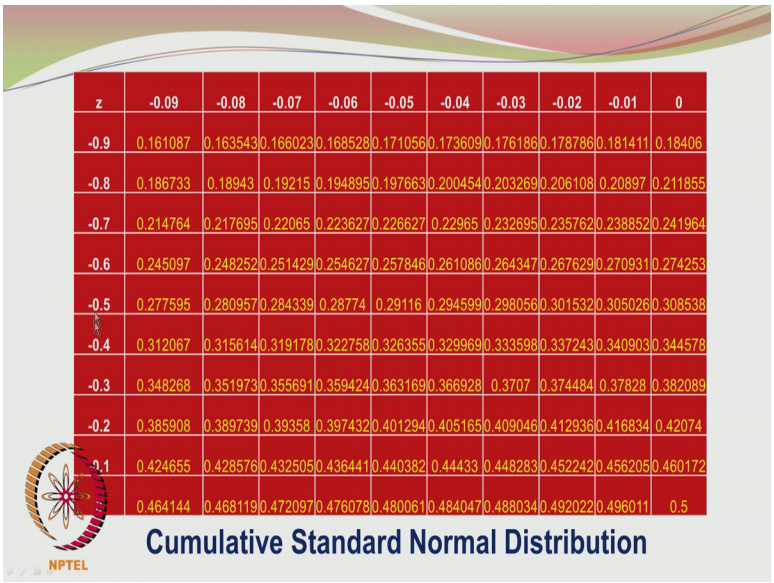
Suppose you want to find the probability-3.55 okay that is 3 decimal points, you locate -3.5 here go to you are right, you will see that -3.555 lies between -3.55 and -3.56, so you may want to interpolate between these 2 values. Even up to 2 decimal places the chart is pretty useful, but suppose you want to find the beyond 2 decimal places, then you have to do some interpolation, the values are likely to become slightly erroneous at the third decimal or fourth decimal, which is okay for most practical purposes.

If you want very accurate values of probabilities, then you have to resort to spreadsheet or any statistical analysis software.

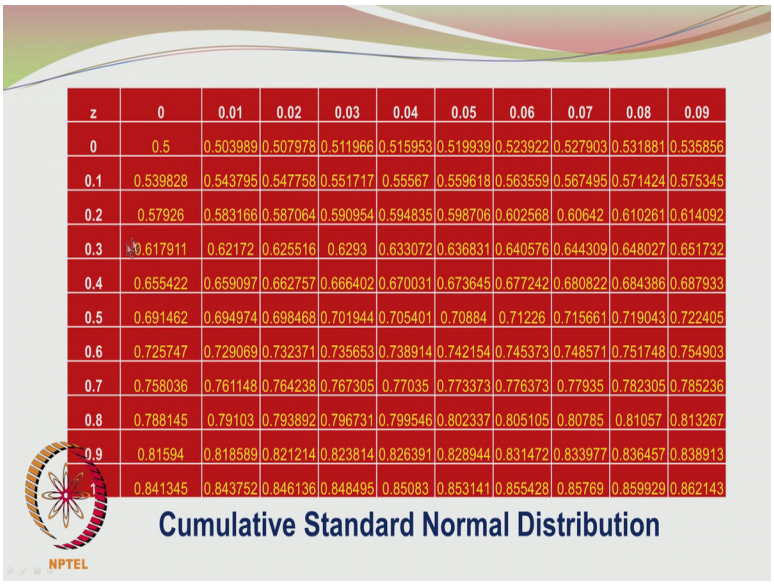
(Refer Slide Time: 27:32)



(Refer Slide Time: 27:43)



(Refer Slide Time: 27:47)



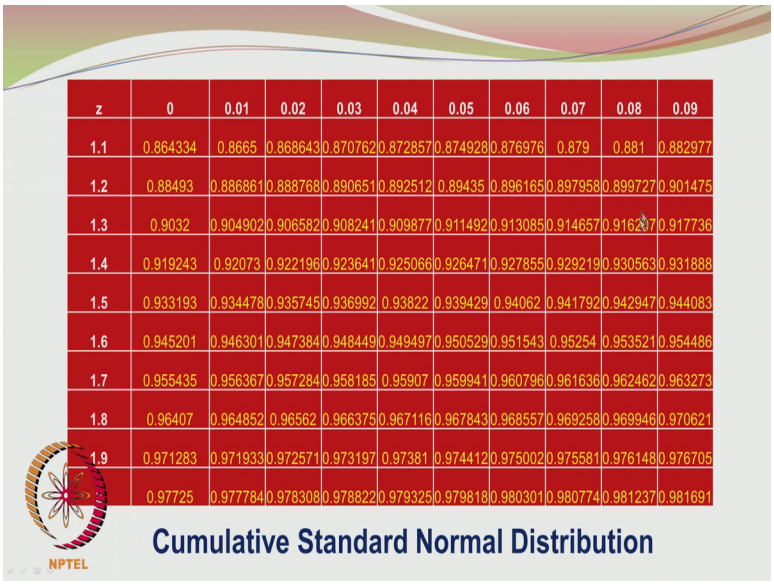
z	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0	0.5	0.503989	0.507978	0.511966	0.515953	0.519939	0.523922	0.527903	0.531881	0.535856
0.1	0.539828	0.543795	0.547758	0.551717	0.55567	0.559618	0.563559	0.567495	0.571424	0.575345
0.2	0.57926	0.583166	0.587064	0.590954	0.594835	0.598706	0.602568	0.60642	0.610261	0.614092
0.3	0.617911	0.62172	0.625516	0.6293	0.633072	0.636831	0.640576	0.644309	0.648027	0.651732
0.4	0.655422	0.659097	0.662757	0.666402	0.670031	0.673645	0.677242	0.680822	0.684386	0.687933
0.5	0.691462	0.694974	0.698468	0.701944	0.705401	0.70884	0.71226	0.715661	0.719043	0.722405
0.6	0.725747	0.729069	0.732371	0.735653	0.738914	0.742154	0.745373	0.748571	0.751748	0.754903
0.7	0.758036	0.761148	0.764238	0.767305	0.77035	0.773373	0.776373	0.77935	0.782305	0.785236
0.8	0.788145	0.79103	0.793892	0.796731	0.799546	0.802337	0.805105	0.80785	0.81057	0.813267
0.9	0.81594	0.818589	0.821214	0.823814	0.826391	0.828944	0.831472	0.833977	0.836457	0.838913
1	0.841345	0.843752	0.846136	0.848495	0.85083	0.853141	0.855428	0.85769	0.859929	0.862143

Cumulative Standard Normal Distribution

So I have covered the broad range, so we have in this particular table from -2.9 to -2 and so on until I get 0, and from 0 again I start from 0.1, 0.2 so on. I want to find the normal probability value corresponding to let us say 0.44, I hit 0.4 here then go horizontally to my right until I reach 0.44, I read out the probability 0.67. This means that probability of the random variable lying taking values below 0.44 is 0.67.

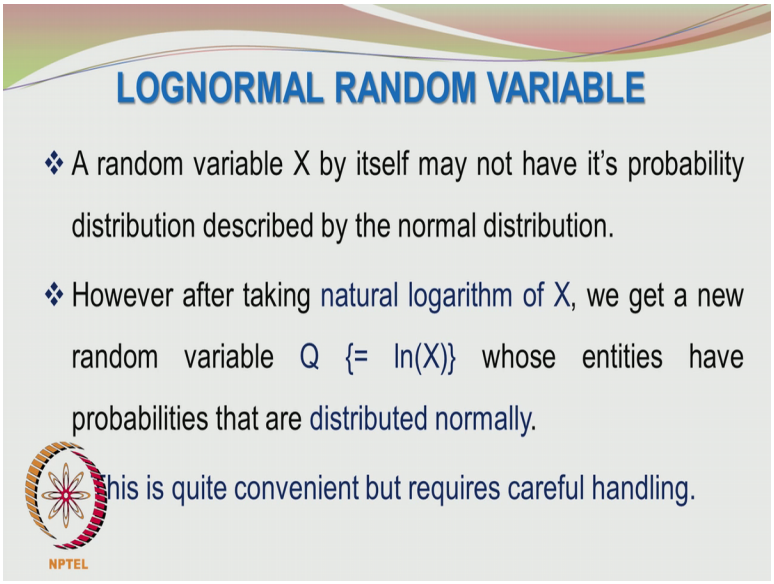
Since I have crossed the origin, I have crossed the area under the curve of 0.5 and so the values would be higher than 0.5 at z=0, you see the probability value is 0.5, probability of the random variable z taking values below 0 that is -infinity to 0=0.25, you are describing the left half portion of the curve. So even at 1.09 you are covering up to 86% of the area under the curve, the probability of a random variable the standard normal random variable capital Z taking a value 1.09 or lower is 0.86.

(Refer Slide Time: 29:23)



If you go further down, you see that you have reached 0.98 at z value of 2.09.

(Refer Slide Time: 29:32)



Now we come to the lognormal random variable, I told you at the beginning of the lecture that even if the original random variable X is not following the normal distribution, in some cases if you make a simple transformation from x to ln X, then it starts to behave in a normal fashion okay. I am not implying that the random variable X the original random variable X was behaving abnormally previously, and once it was converted to ln X it starts behaving normally.

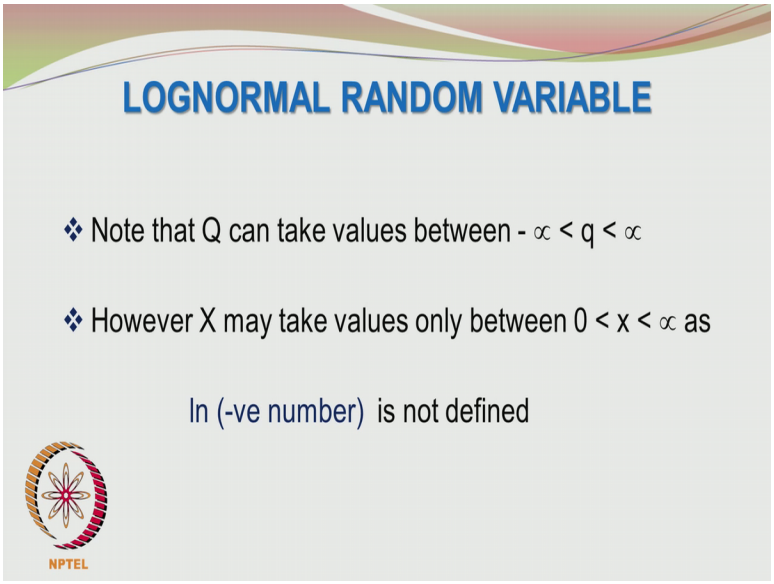
I really mean that it was following some other probability distribution in its original form, when it was going by random variable X but once you have converted it to ln X, the probability

distribution is the normal distribution, it is not going to happen for all cases, so you have to be a bit careful here, you make the conversion from x to ln X, and see whether the distribution has a bell shaped curve or a normal like curve.

So the next thing is let us defined the random variables Q as ln of X, please remember that when use subject a random variable to a mathematical transformation of any kind okay, the transformed variable is also a random variable. For example, you have x and then you convert it from x to x+2 by adding 2, you define a new random variable as x+2, this x+2 let us called it as y, y is also considered to be a random variable.

Similarly, when you have a random variable X, you subject it into a transformation and make it into ln X, then the random variables q which=ln X also has a probability distribution.

(Refer Slide Time: 32:11)

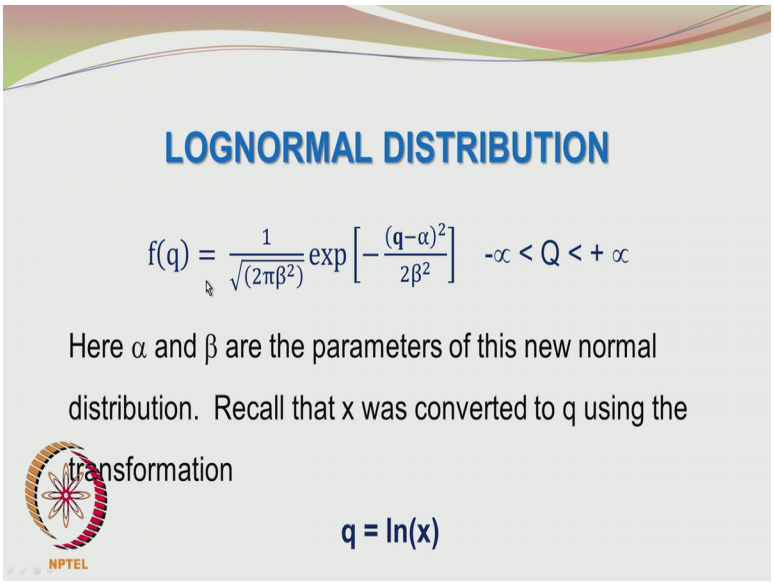


When the random variable has been converted to ln X, the range would be from -infinity to +infinity. However, when the random variable was in the original primitive form X, the range must be only from 0 to +infinity only positive values are allowed okay. The reason is if the random variable X had a negative value, when you take ln of a negative value it is undefined okay. So you cannot really have X values that are negative, this is not a restriction.

In many physical cases you have only positive values that are possible for the random variable, for example particle size distributions, you can have very small particle sizes of 10 power-3 or 10 power-4 meters and so on in the micron range or sub-micron range or even the nano range, but they are all positive. When you take the ln of number which is <1, then it will become negative and so in the transformed domain the range can be from -infinity to +infinity.

So if you look at this particular slide, the transform the random variables Q may take values between -infinity<q<infinity. The primitive random variable X may take values only between 0 to infinity, it cannot take a valuable lower than 0, because ln of a negative number is not defined.

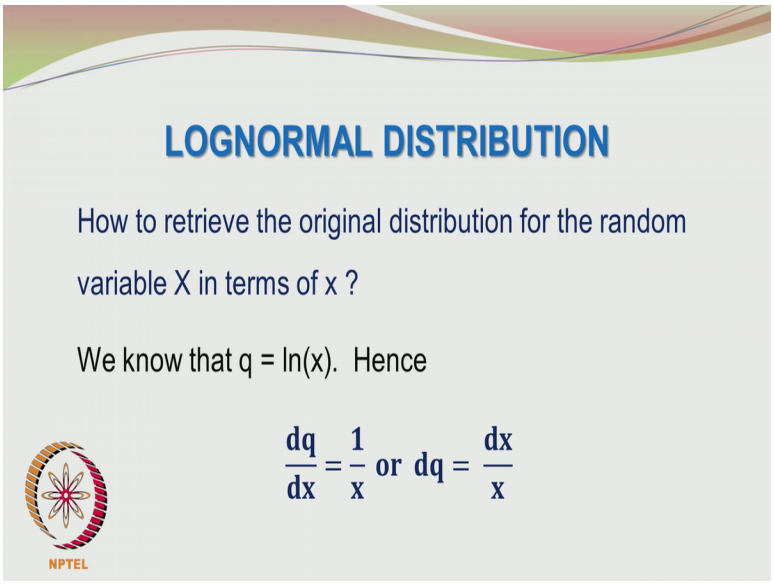
(Refer Slide Time: 34:25)



Let us now look at the form of the lognormal distribution, it is not something new since it is going to follow the normal distribution, we have f of q=1/root 2 pi beta square exponential-q-alpha whole square/2 beta square, the value is -infinity as the lower limit and +infinity as the upper limit. The parameters are alpha and beta, we use q here, q was obtained by taking the natural logarithm of x okay, this is an important point, please do not put x here.

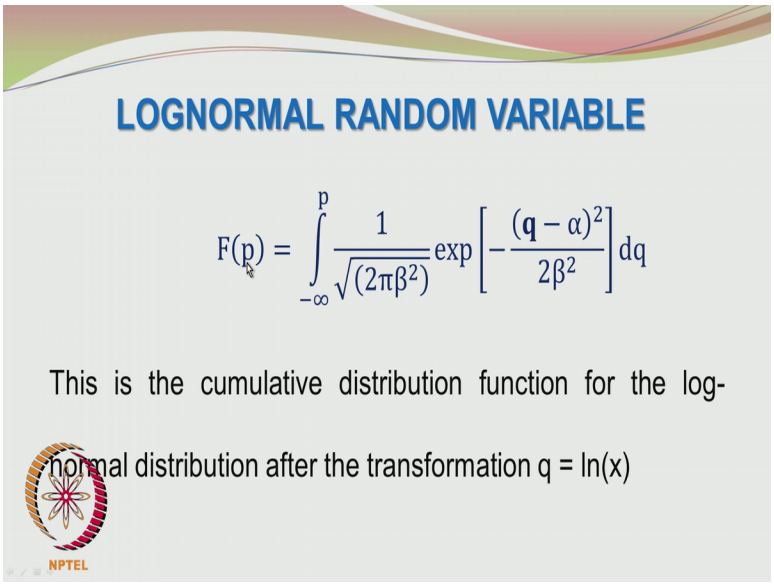
You have to convert it into q by taking the natural logarithm of x, then use q here, then you can use the normal distribution characteristics.

(Refer Slide Time: 35:28)



So you have the random variable X, and you also have the transformed variable Q, so after the transformation you got a normal distribution, what was the distribution like in the original form? How was the probability distribution defined in terms of the original random variable x that is very interesting, we can find it pretty easily? We know that q =ln x, so we differentiate q with respect to x, we get dq/dx=1/x or dq=dx/x. This we can use to retrieve the original form of the distribution in terms of x.

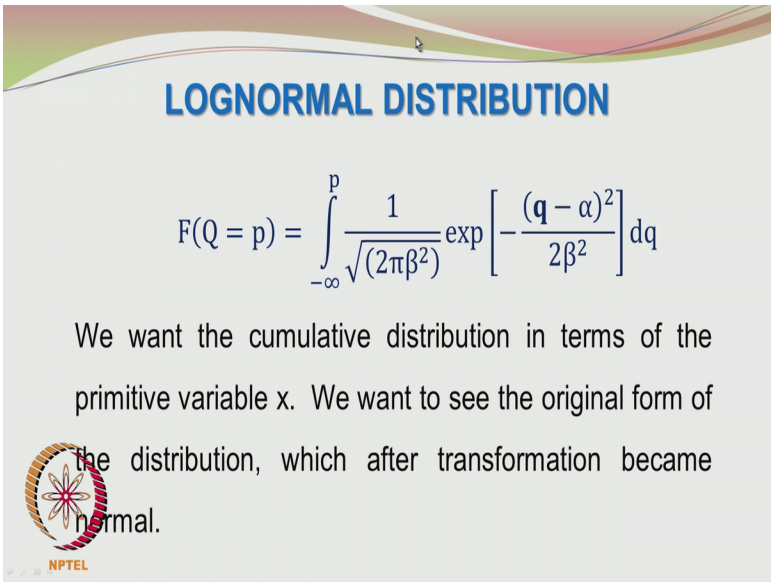
(Refer Slide Time: 36:21)



The cumulative distribution function for the lognormal random variable is given by f of p= -infinity to p 1/root 2 pi beta square exponential-q-alpha whole square/2 beta square dq, this is nothing but the cumulative distribution function, we have already seen this in one of our earlier

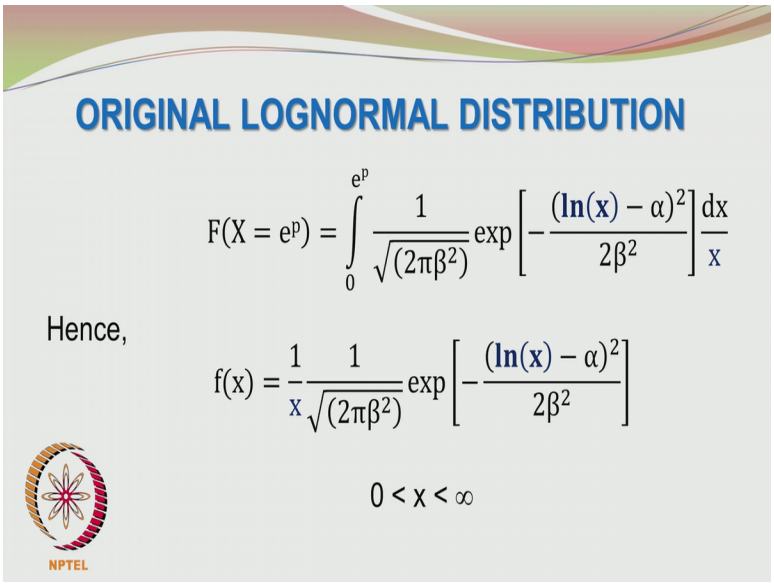
slides. So to find the value of the probability <=p, we do the integration up to p, now we can use this find the original form of the probability distribution expressed in terms of x, here we are using ln x or q.

(Refer Slide Time: 37:05)



Here, we put instead of q we put ln x, instead of dq we use a substitution dx/x. When we convert q into x, we have to make sure that the limit are also appropriately changed.

(Refer Slide Time: 37:31)

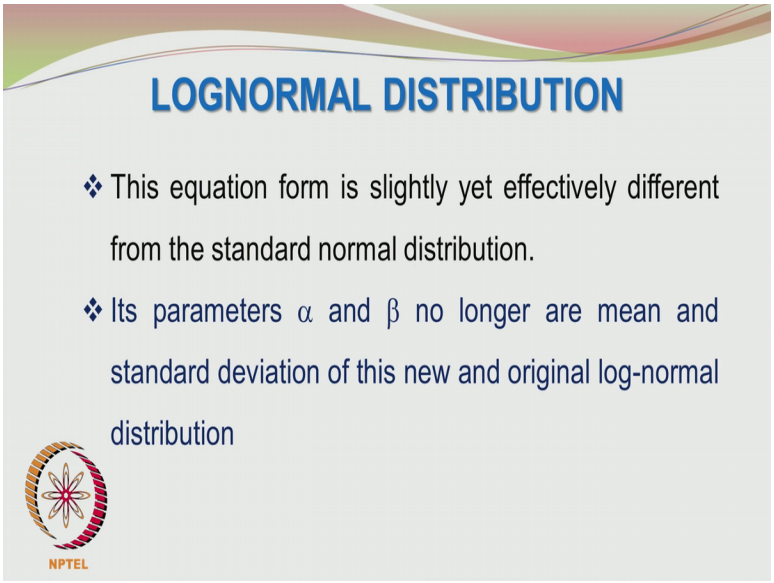


When we do that we see that the lower limit of q which was -infinity has become 0, when we converted into the x domain. And the value of p became e power p, when you converted into the x domain, and instead of q we have put ln x, and instead of dq we put dx/x. And this represents

the cumulative distribution function in terms of the original random variable x. So the probability density function is given by 1/x*1/root 2 pi beta square exponential-ln x-alpha whole square/2 beta square.

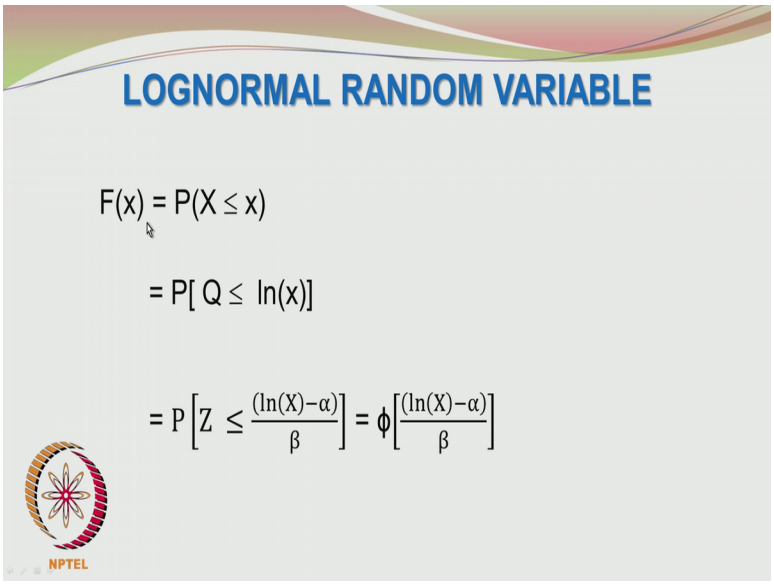
And the value of x is between 0 to infinity, this is very interesting here this f of x looks quite similar to the normal distribution, but note that instead of x we have used ln x, and there is an additional 1/x term here, even though the differences are seemingly slight they are quite significant.

(Refer Slide Time: 39:02)



Important thing to note here in this distribution alpha and beta are not the mean and variance or rather standard deviation of this distribution. For the normal distribution mu and sigma actually represented the mean and standard deviation, but for this distribution the lognormal distribution alpha and beta do not represent the mean and standard deviation, this is something which we have to remember.

(Refer Slide Time: 39:45)

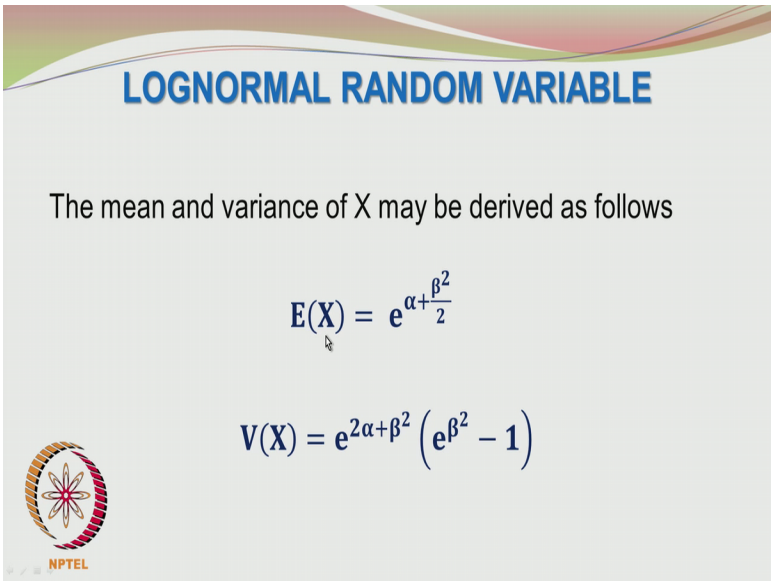


This is the cumulative distribution function by now it should be familiar to u, F of x=probability of X<=x, and that maybe written as probability of Q<=ln of x. How did you get this? You have to just take ln of x ln of small x, and so ln of capital X became Q<=ln x. To find the probabilities we have to convert them into the standard normal form that is pretty easy, what we do is we subtract alpha from ln x and divide by beta.

Please note that alpha and beta are the mean and standard deviation of the distribution in the transformed case, you converted x to ln x, and that started behaving in a normal fashion, and so the parameters alpha and beta did represent the mean and standard deviation of that normal distribution provided the transformation x to ln x had taken place okay, this is very important. And then you can treat it as a normal distribution as like any other normal distribution and convert it into the standard form.

But if you are not using ln x, but you are using x directly then you will have to use this form of the probability density function, and there alpha and beta cannot be interpreted as mean and standard deviation. So you have the standard form ln x-alpha/beta, probability of Z<=ln of x-alpha/beta may be represented as the cumulative distribution function phi of ln of x-alpha/beta.

(Refer Slide Time: 42:00)



I told you that in the original form you may recall me telling you that alpha and beta are not the mean and standard deviation of the lognormal distribution, so how do we find the mean and the standard deviation of the lognormal distribution? It is quite simple, we use the parameters alpha and beta the expected value of X=mu, and that is given in terms of e power alpha +beta square/2, and the variance of x is sigma squared okay, and that is given in terms of e power 2 alpha +beta square*e power beta square-1.

So to find the mean use this formula, to find the variance use this formula. So we have covered 2 important distributions in this lecture. The first one was the normal distribution, and the next one was the slightly confusing lognormal distribution. But after solving a few problems, you will have no such conclusion. We will take some illustrative problems and solve them using the normal probability tables, and the concepts will become clear.