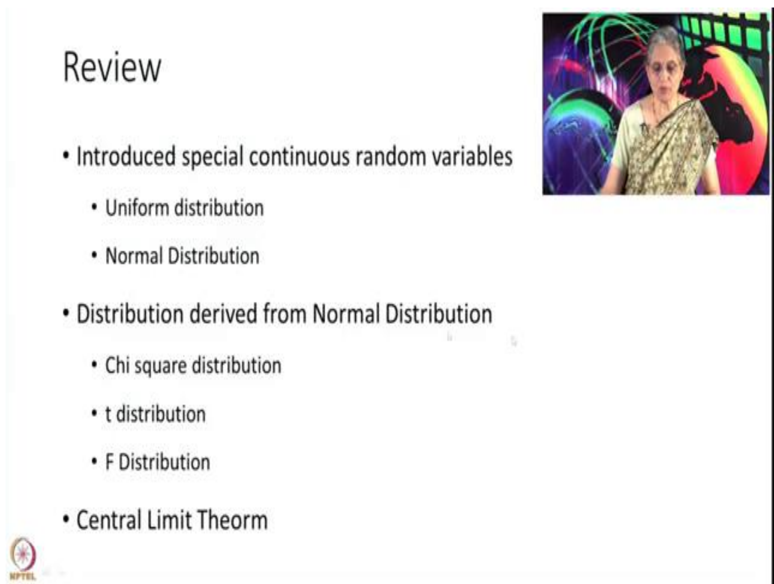
Dealing with Materials Data: Collection, Analysis and Interpretation
Professor. Hina A. Gokhale
Department of Metallurgical Engineering and Materials Science
Indian Institute of Technology, Bombay
Lecture 35
Special Random Variables 5

Hello and welcome to the course on Dealing with Materials Data. In the present sessions, this is the fifth one in the series of this sessions on special random variables. We are talking about certain special distribution functions, which we encounter very commonly in our day to day experimental work.

(Refer Slide Time: 00:45)



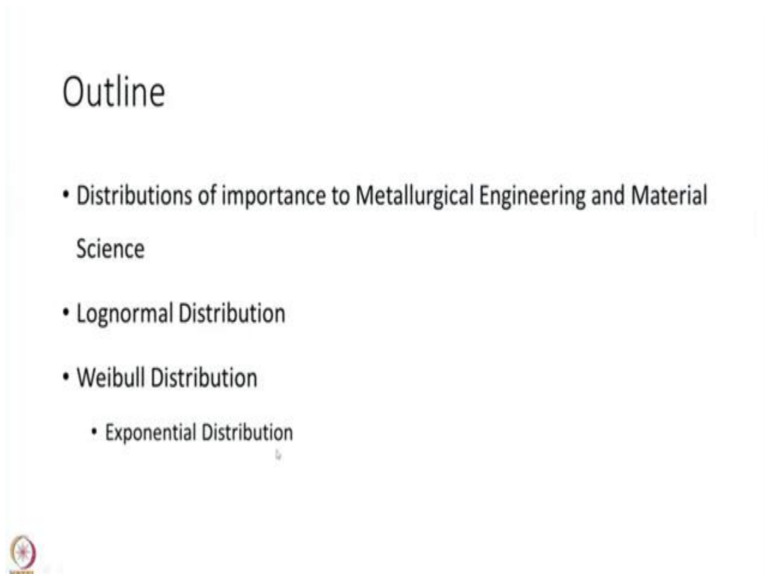
Let us quickly review, we first introduce some continuous random variable before that we introduce some discrete random variables, uniform, discrete uniform random variables then Bernoulli trials and three distribute derived out of Bernoulli trials, we gave a 3D APFIM example there, then we introduce the uniform continuous random variables, and we introduce uniform random, uniform random variable or Uniform Distribution.

Then we talked about in details on Normal distribution. We also discuss some distributions which are derived out of Normal distribution, they do not largely naturally occur. However, they are useful to carry out certain inference on the random variable which originated from the Normal distribution and the last session, we discussed about the Central Limit Theorem.

And that basically showed that how Normal distribution is kind of all pervading distribution primarily because it said that with the condition that it has a finite variance, a large number of random variables coming from, independent random variables coming from identical distribution.

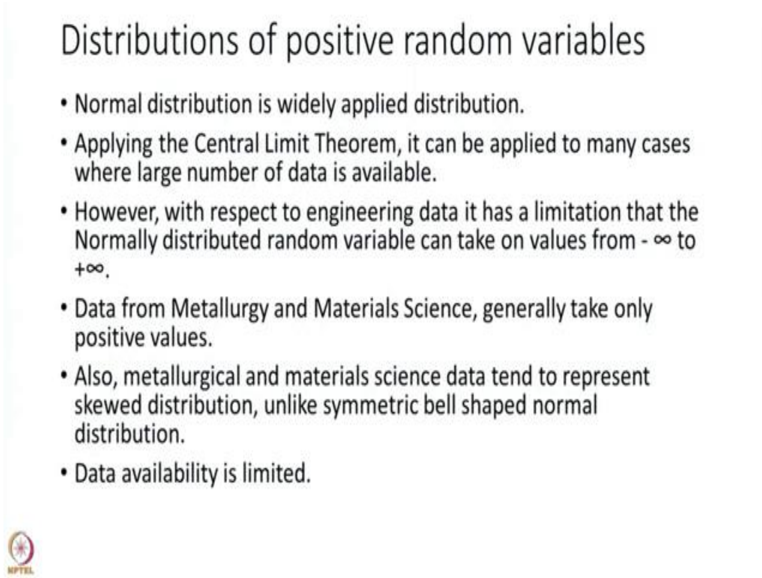
And if you take its mean value then as n tends to infinity roughly we can say that mere the mean value tends to Normal distribution with, the average value of the random variable follows the Normal distribution with the common mean and common sigma square divided by n or the random, the average minus the mean divided by its variance follows a standard normal variable that is a correct way to put it.

(Refer Slide Time: 02:59)



Now, we are going to come to some of the distributions which are of importance to metallurgical and materials engineering and that is very specific, but in general it is good for the engineering distributions or engineering data. The two distribution that we wish to introduce right now is Lognormal Distribution and Weibull Distribution and then we will follow up with the Exponential Distribution, which is a special case.

(Refer Slide Time: 03:28)



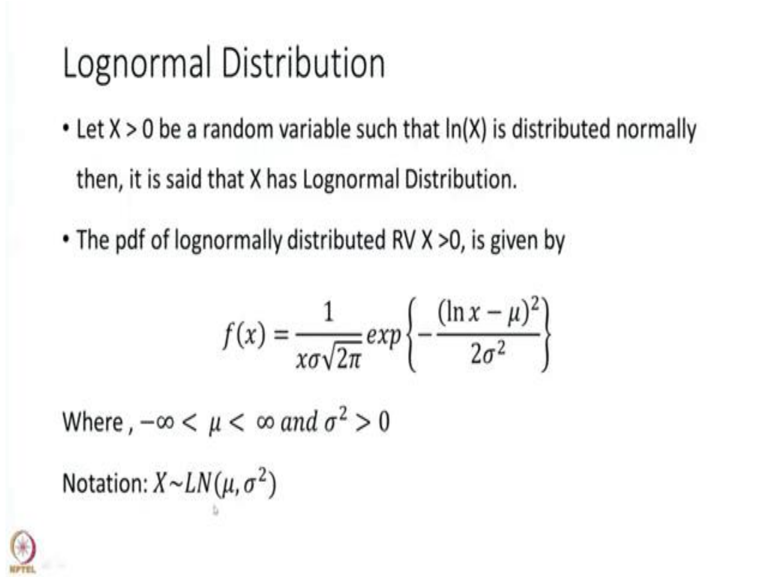
Why this special treatment to some of this distribution, particularly coming from a metallurgical engineering data? Well, normal distribution is a widely applied distribution. Central theorem makes it almost all pervading distribution. However, it has one limitation with respect to engineering data. In Normal distribution the random variable varies between minus infinity and plus infinity. It means that it can take any value from the realign, while data from metallurgy and material science generally take on only positive values.

Now, you may assume the distribution to be normal and carry out your inference. Sometimes you encounter a very absurd result and it may amount to the fact that you have assumed normality which takes on all kinds of values from negative infinity to positive infinity and therefore, it creates this absurdity.

Second thing is that this metallurgical and material science data in particular it has been seen that they have they are generally a skewed data, they are not beautifully symmetric bell shaped histogram, they do not represent those such beautiful histogram they generally represent the skewed distribution also to apply the Central Limit Theorem as we say if the distributions are skewed.You need at least 30 data points or at times you may need even more than that if the skewness is very high in that case, such a data availability in material science and in particularly in metallurgical engineering is very rare, it is very expensive to generate some data.

And therefore, it is important that we learn about certain distributions, which are defined over the positive part of the straight line that is for the positive random variables. And we will also see that how they have been used and applied for in the area of metallurgy and material science.

(Refer Slide Time: 06:05)

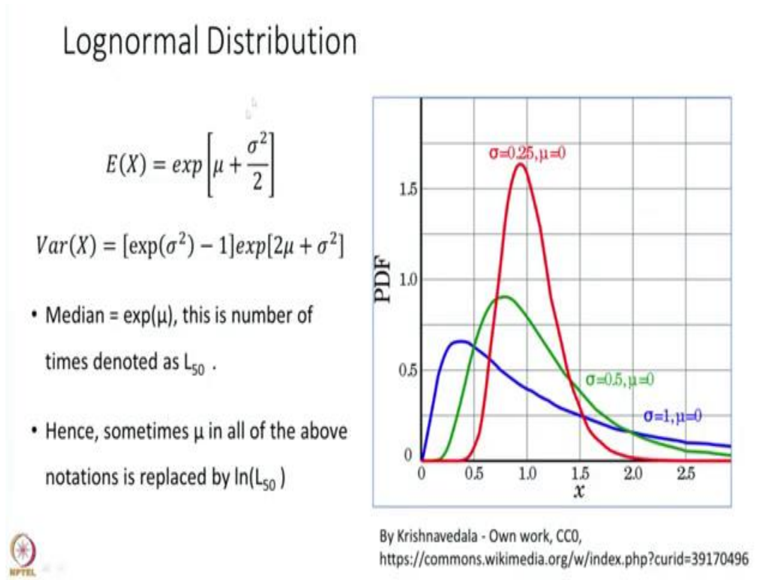


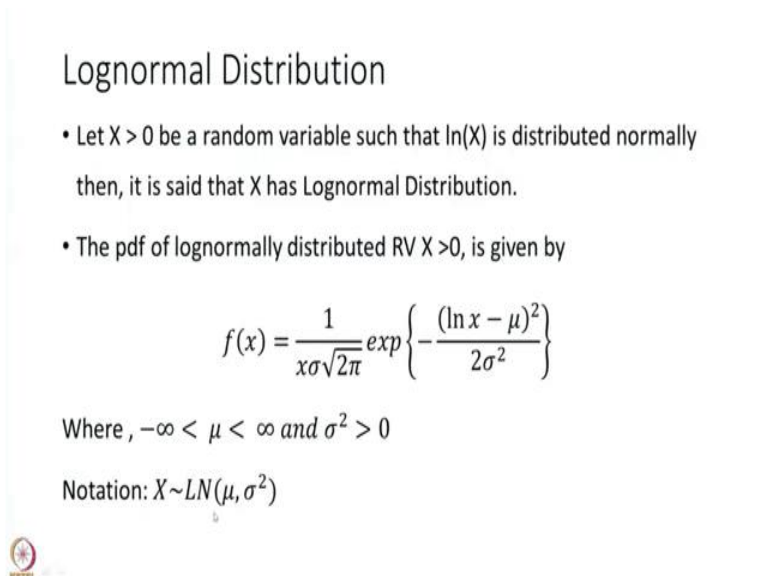
So, first we introduce what is known as Lognormal Distribution, as the name says, it is a positive random variable, which has the natural log of that random variable follows or normal distribution then you say that X has a Lognormal distribution. So, there is a random variable whose natural log follows Normal distribution then we say that X follows log Normal distribution, such a distributions PDF is given in this manner,

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} exp\left\{-\frac{(ln\, x - \mu)^2}{2\sigma^2}\right\}$$

Here also the two parameters are there one is a mu and the other is sigma, mu is a mean value of ln(x) and sigma square is a variance of ln(x). Therefore, mu varies between minus infinity and plus infinity and sigma square is always positive. The notation we use is X is distributed $X \sim LN(\mu,\ \sigma^2)$, there are two parameters to it.

(Refer Slide Time: 07:26)





Here I have shown you some of the plots I have picked it up from the Wikipedia. The reference is given here. Here, what I want to show is that the mean value is kept at 0 and the variance of the normal part of the distribution is varied from 1 to 0.5 to 0.25 and you can say that with the change in sigma value, remember that this is not sigma square, this is sigma.

So, if you try to regenerate it, please make sure that it is with sigma and not sigma square given here and then you can see that as the sigma changes, the shape of the curve changes, as the sigma becomes smaller, the curve becomes more and more skewed. And also please notice that this is always skewed towards the right. Lognormal distributions are always skewed towards right as sigma, you can see that when sigma is 0.25, it comes very, I think I made a mistake in my first statement.As the value of sigma increases, the skewness decreases. When sigma is large, the skewness is large and when the sigma is small, the skewness decreases and you can see that here at sigma equal to 0.25 this is almost like a normal distribution curve, but there is you can see the tail part of it and you can make out that it is still a skewed distribution.

The expected value is given $E(X) = exp\left[\mu + \frac{\sigma^2}{2}\right]$

thus I leave it for you to derive it and the variance is expressed in this manner

$$Var(X) = [\exp(\sigma^2) - 1]exp[2\mu + \sigma^2]$$

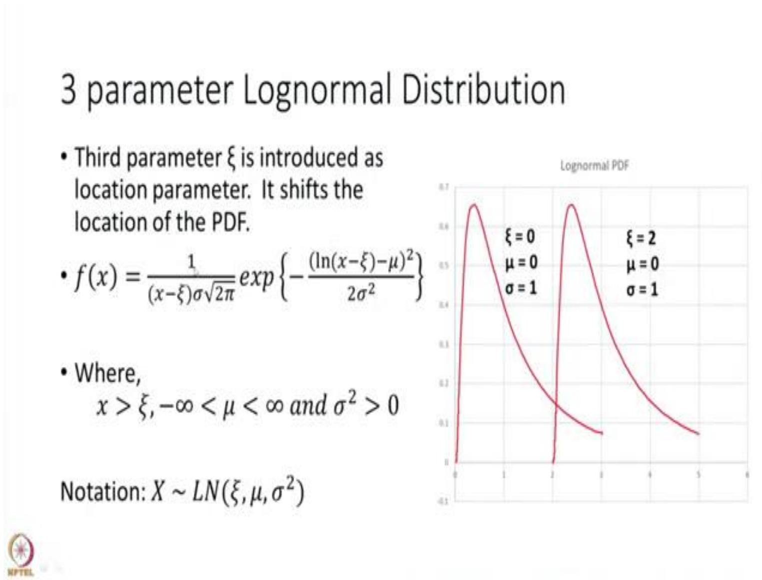
$$Median = exp(\mu)$$

So, exponential of mu is the, here it is 1. So, this is the median value of the distribution, for every distribution this is the median value.

So, this is all centered around the median value and median value a number of times is also denoted by L sub 50 and sometimes mu in the, all the above equation these equations as well as these equation is replaced by an ln L50. Because remember that L50 is exponential of mu therefore, mu is lognormal of it is a natural log of L50.

Now, please remember here, let us come back that sigma square decides the shape of the distribution and therefore, here sigma is called the shape parameter. Unlike, Normal distribution, it is not a scale parameter it is a shape parameter.

(Refer Slide Time: 10:45)



Now, if you want to introduce location parameter, then it is a third parameter and it makes it a three parameter Lognormal distribution.

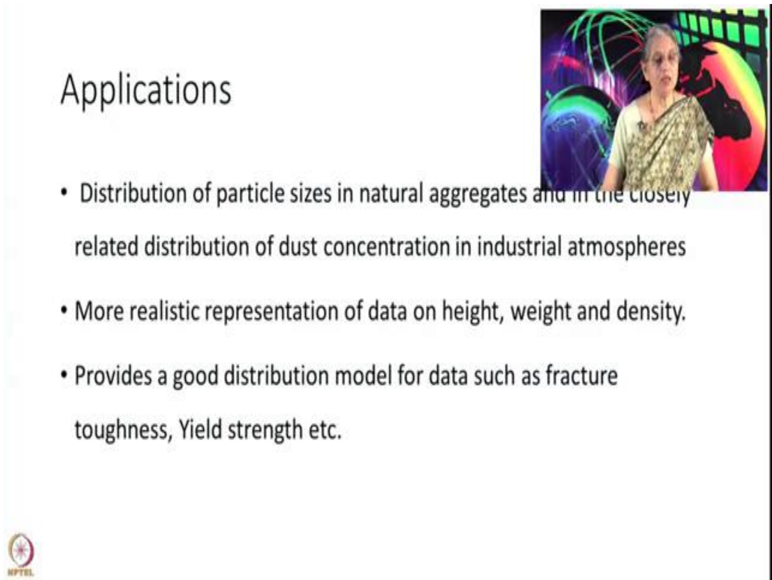
Where $X = (x - \xi)$ where your assumption now is not x greater than 0, but x is greater than psi,

$$x > \xi,\ -\infty < \mu < \infty \text{ AND } \sigma^2 > 0$$

And then the form of the distribution is this, it is denoted as x is distributed as : $X \sim LN(\xi,\ \mu,\ \sigma^2)$.

Let us look at some of this PDF plot. Here I have kept to show the effect of the location parameter which is psi. The first graph shows that psi is 0. You see this is when the location is, location is 0 and you see that this is where the x values are larger than 0. So, it starts from 0, 0 and it goes up and down. When you make psi is equal to 2, the whole graph is sort of shifted by 2 units and therefore, because starts from 2 and exactly the same trend it follows. So, that is why it is called a location parameter because it shifts the location of the plot or of the PDF.

(Refer Slide Time: 12:22)

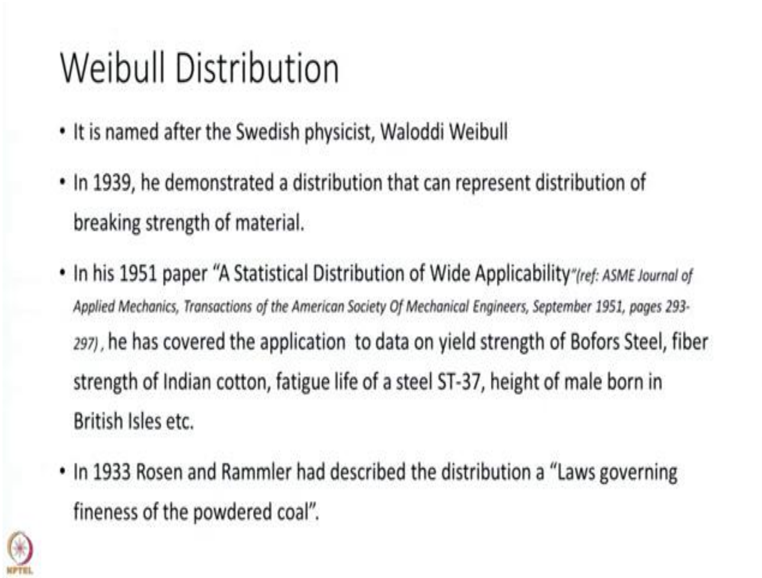


This Lognormal distribution is most commonly used to estimate the particle size distribution. So, another way is also that it is heavily used sometimes to represent the data such as the height, weight of a human being, because, generally height and weight is the modeled as a normal distribution.

But as I said height and weight cannot take a negative value at all. And therefore, it is actually a positive distribution. It is a positive random variable and therefore, it is number of times estimated the distribution of height and weight is estimated very well with the Lognormal distribution.

Also, I find that the density is also very well modeled as a Lognormal distribution. It also gives a good model to express the distribution of fracture toughness, yield strengths, etcetera, etcetera. So, it has a very typical, very, very specific and many applications in the field of metallurgy and material science.

(Refer Slide Time: 13:46)



Now, we move on to the next distribution, which is a Weibull Distribution. Weibull Distribution is named after a Swedish physicist, Waloddi Weibull. In 1939, he first time demonstrated that a distribution that can represent the distribution of breaking strength of material and he must have got so excited with his discovery or this particular finding of the distribution.

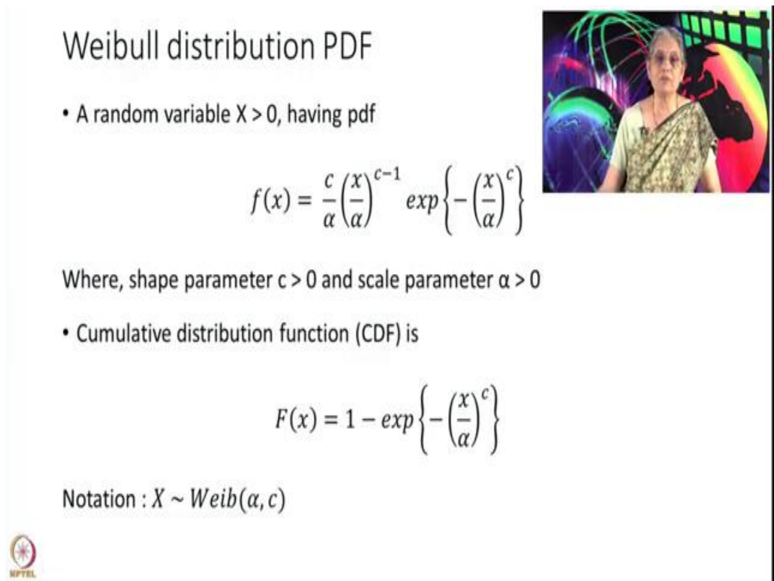
That he wrote a paper in 1951, which titles a Statistical Distribution of Wide Applicability. I have given the reference and I strongly recommend that everyone should go through this paper, because it is a very interesting account of how the physical phenomena can be looked into and can be used to derive the statistical distribution. It is a very beautiful distribution, about two paragraphs of it, but very, very inspiring and very interesting.

In this paper, he said that not just the breaking strength of the material in terms of yield strength Bofors Steel or fatigue of another steel ST-37. He even found that the fiber strength of the Indian cotton, height of the male born in British Isles, he also found certain length of a certain beans which were produced in a certain farm.

All of them tend to follow Weibull distribution. It appears that Rosin and Rambler had in 1933 had also come up with a very similar distribution in connection with particle size distribution. However, the distribution became very more known or more popular, when

Weibull put it forward in his 1951 paper and it has been heavily referred, I again once again I said that I strongly suggest that you please go through this, I have given the full reference of this paper, please go through this paper. It is a very interesting account of how the distribution is derived.

(Refer Slide Time: 16:06)



So, if X is a positive random variable, and it has a PDF that looks like this,

A random variable $X > 0$, having pdf

$$f(x) = \frac{c}{\alpha}\left(\frac{x}{\alpha}\right)^{c-1} exp\left\{-\left(\frac{x}{\alpha}\right)^{c}\right\}$$

Where, shape parameter $c > 0$ and scale parameter $\alpha > 0$

then it is safe to follow Weibull distribution.

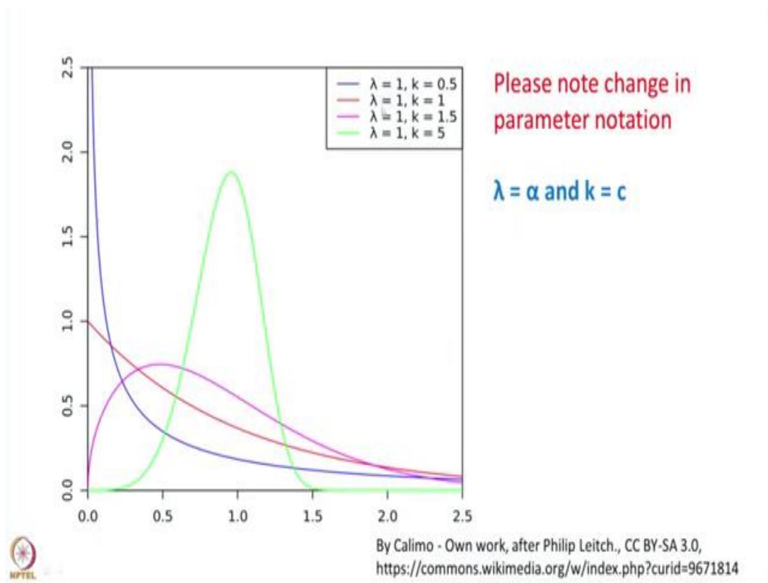
Here we have two parameters, c and alpha, c is called a shape parameter and alpha is called a scale parameter and the cumulative distribution function as much easier expression which is given here, Cumulative distribution function (CDF) is

$$F(x) = 1 - exp\left\{-\left(\frac{x}{\alpha}\right)^{c}\right\}$$

Notation : $X \sim Weib(\alpha,\ c)$

First comes the scale parameter and then comes the shape parameter. Please look at this particular expression and note that if you take double log of this CDF, you will have a linear relationship between c and rather log c and alpha and c and log alpha and this will be one of the ways of estimating the parameters when we come to that. So, we will discuss it in more details at that time.

(Refer Slide Time: 17:35)



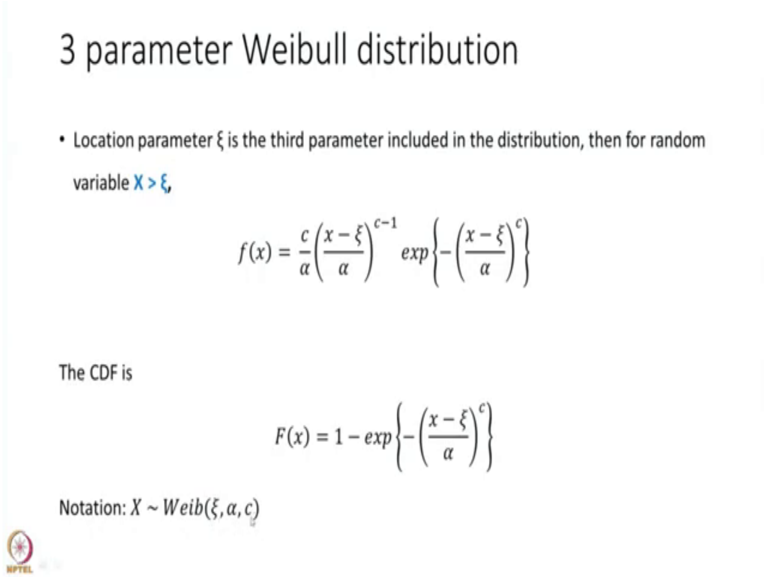
Again I have taken some plots of Weibull distribution from Wikipedia, please remember here, the notation is slightly different, the scale parameter is given as lambda and the shape parameter is given as k. And here again for different shape parameters, the distributions are shown and please note, that this distribution is not just skewed, but it can be skewed in either direction.

That is the beauty of Weibull distribution, it can be skewed in either direction. So here you see a kind of an exponential graph here, you see another distribution which is very typically right skewed distribution. This one, this is for shape parameter is equal to 1.5. This is another skewed, right skewed distribution.

However, when you take k is equal to 5, as such, it looks like a symmetric curve, but notice that tails and you realize that now it is becoming slowly the left skewed distribution, the longer tail is coming on the left side. That is the beauty of Weibull distribution.

In Lognormal distribution only if the data is positively skewed, you can fit data, but Weibull gives you an open field that any data which may be skewed or may have an exponential kind of expression can also be covered as a Weibull distribution, can also be modeled as a Weibull distribution.

(Refer Slide Time: 19:16)



Weibull distribution also has its variant called three parameter Weibull distribution in this exactly the same manner as in the Lognormal distribution. So, there is a location parameter.

Location parameter ξ is the third parameter, then for random variable $\boldsymbol{X > \xi}$,

$$f(x) = \frac{c}{\alpha}\left(\frac{x-\xi}{\alpha}\right)^{c-1} exp\left\{-\left(\frac{x-\xi}{\alpha}\right)^{c}\right\}$$

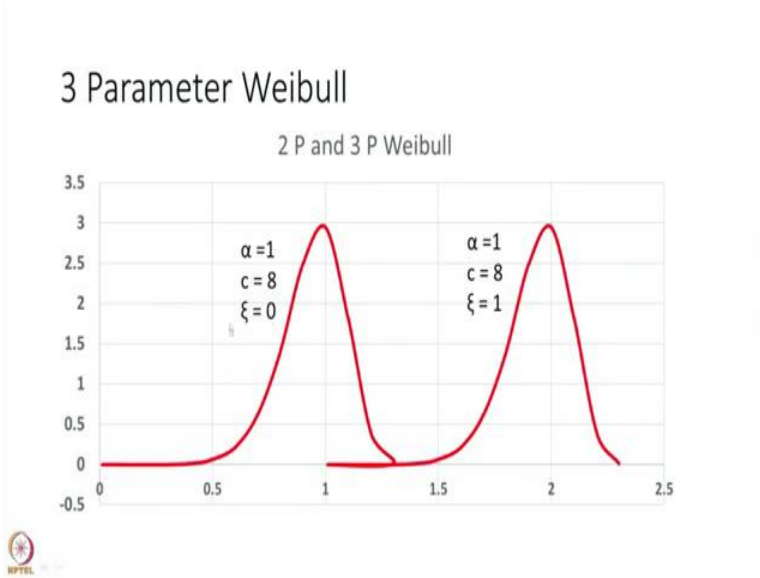
The CDF is

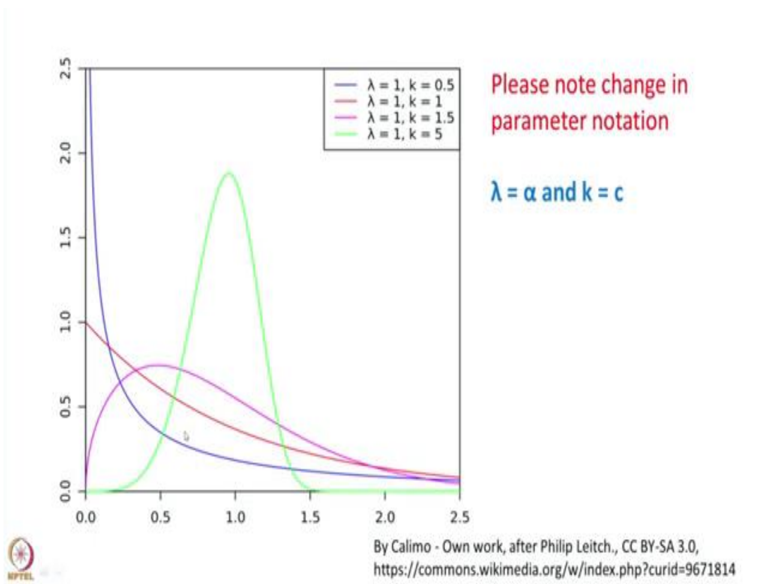
$$F(x) = 1 - exp\left\{-\left(\frac{x-\xi}{\alpha}\right)^{c}\right\}$$

Notation: $X \sim Weib(\xi,\ \alpha,\ c)$

This is the expression of PDF you please recall that the expression only changes that x has been replaced by x minus psi in all the places and the CDF is also generated in the same way and in this situation the notation is X is distributed as Weibull psi alpha, c. It means that location, scale and shape.

(Refer Slide Time: 20:14)

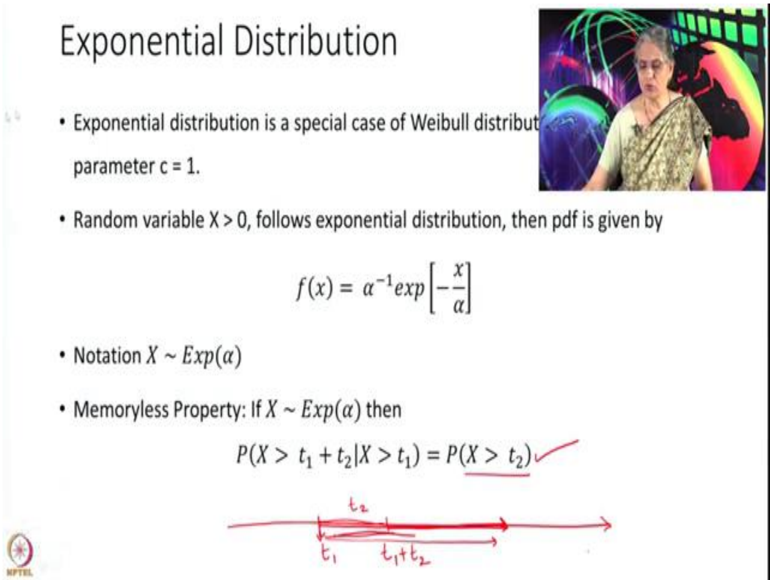




How does the three parameter Weibull differ? This is a two parameter Weibull because I have a psi is equal to 0 and here I have taken psi equal to 1, you can see that since I have taken the shape parameter to be 8, it is left skewed distribution and it simply shifts is location from instead of starting from 0, I can say that x greater than 0. So, x starts from 0 onwards, here it starts 1 onwards, otherwise it has simply shifted.

I also want to tell you from the graph, that as the value of shape parameter increases, the shape changes from the right skewness to the left skewness. So, here also, we have very specifically showed a left to distribution arising out of Weibull and how it shifts, it is actually a location it shifts the location of the distribution.

(Refer Slide Time: 21:29)



Now, there is one distribution, Exponential distribution, which is very special case of Weibull distribution, where the shape parameter is 1, you please recall shape parameter 1, the red color distribution is represented is a special distribution. It is called Exponential Distribution.

Exponential distribution is a special case of Weibull distribution, where shape parameter c = 1.

Random variable X > 0, follows exponential distribution, then pdf is given by

$$f(x) = \alpha^{-1} exp\left[-\frac{x}{\alpha}\right]$$

Notation $X \sim Exp(\alpha)$

I have purposefully left it out please derive the CDF of Exponential distribution, it will be a good mathematical calculating algebraic exercise for you.

Also please note that before going any further to the memory less property, you can have a three or rather two parameter Exponential distribution, here you have only one parameter which is alpha. Now, you can introduce a psi which is greater than 0 and say that x is greater than psi, follows a three parameter exponential distribution where you will replace x by x minus psi. So, you can generate a three parameter exponential, two parameter exponential distribution as I stand corrected and then you will have x distributed as exponential.

First you will have psi and then you will have alpha. This distribution has a wonderful property called a memory less property. What does it mean?

Memoryless Property: If $X \sim Exp(\alpha)$ then

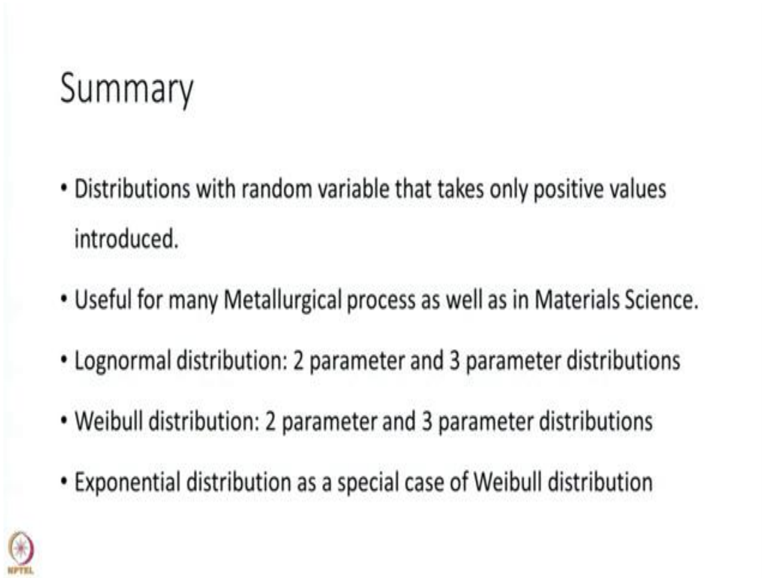
$$P(X > \ t_1 + t_2 | X > t_1) = P(X > \ t_2)$$

This is a very important and very interesting property, let us understand this.

Suppose this is a timeline, this is a time the t1 and this is a time t1 plus t2, this is the value of t2, what it really says is that you already know that the random variable is larger than this. Then if you try to find out that the random variable is larger than this given that it is already larger than that only means that it is beyond the time t2, what it has done in this range, it has already gone beyond the t1, it has forgotten, it only remembers that it has gone further t2 time.

That is all it says. So, this is called that it has forgotten the memory that it was already larger than t1. But now, it only considers itself as larger than t2 this proof also I leave it to all of you to consider it. It is a very interesting proof you should try it. It is not very difficult.

(Refer Slsside Time: 25:15)



So, with this, let us summarize what we did today. We considered all the distribution with the random variable which takes on only positive values. Why did we do that? Because in material science and metallurgical engineering, most of the data we encounter is positive in nature. You can assume normality and carry out the inference but at times you encounter certain absurdity in your results.

And that is likely to arise because you have considered a distribution which is supposed to have its random variable coming from both positive and negative side, while in reality, you have caught the random variable which is only positive. So, we have introduced specifically distributions which have great applications in the field of metallurgy and material science, and it takes on a positive value.

So, we introduced two distributions of this nature Lognormal distribution, two parameters and three parameter distribution, Weibull distribution also two parameters and three parameter and then Exponential distribution as a special case of Weibull distribution. I would like to let every one of you know that these are not only two distributions useful to with having a positive random variable and useful to metallurgy and material science.

There are many such distributions. One of them is Birnbaum Saunders distribution, there is a gamma distribution, there is beta distribution, there are many such distributions, but right now we have introduced most commonly useful distribution in the field of material science and metallurgical engineering. With this we complete the sessions on introducing the special random variables.

We can have a quick look at it, we started off by saying that there is a random variable which is a function, from the probability space to a real line. And then for such a random variable, we talked about it expected value, its variance etcetera. Then we introduced the special random variables, which take on very special distribution.

So, we talked about the discrete random variables of special kind. Where, we had Uniform distribution, Bernoulli trials, Binomial distribution, Geometric distribution, we also had a Negative Binomial distribution, Hypergeometric distribution and all of this we showed through 3D APFIM and other examples that they are all useful to us in the field of metallurgy and material science, then we came to the Continuous distribution in which we talked about the Continuous Uniform distribution.

And we showed that in random number generation, it plays a very central role, then we introduced Normal distribution, we had a long discussion on Normal distribution, because, through Central Limit Theorem, we saw that it is largely all pervading distribution, if you have a large amount of data, we also got introduced to derivatives of Normal distribution such as Chi squared distribution, t distribution and F distribution.

We have not seen the applications of it directly but it is a preparation for the future, a work that we are going to do in estimation of parameters and testing of hypothesis. Then of course, we talked about the Central Limit Theorem and then we introduce ourselves to the distributions which are useful to the metallurgical engineering and material science because most of the time random variable takes on only positive values.

We introduced the commonly used such distribution as Lognormal distribution, Weibull distribution, and we introduce Exponential distribution which is a memory less property as a special case of Weibull distribution. Thank you.