Mine Automation and Data Analytics

Prof. Radhakanta Koner

Department of Mining Engineering

IIT (ISM) Dhanbad

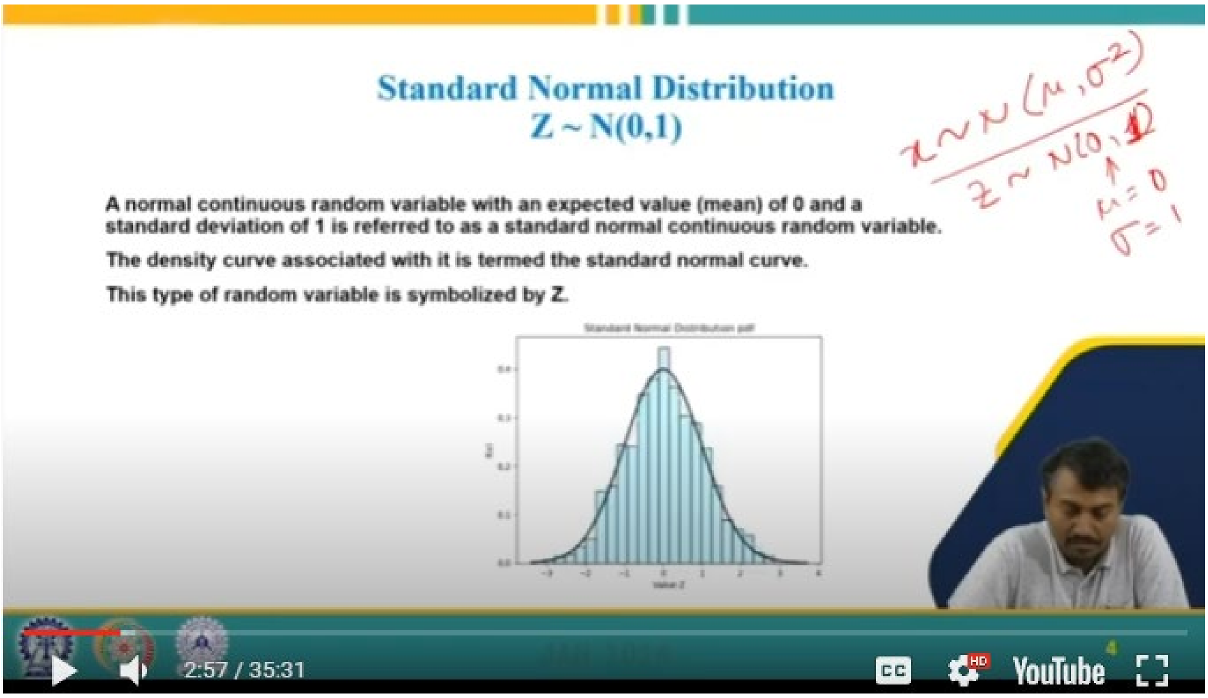
Week - 8

Lecture 40: Continuous Random Variable Part II

Welcome back to my course, mine automation and data analytics. So, in this lesson, we are going to discuss the remaining part of the continuous random variable. In the last lesson, we discussed the continuous random variable, pdf, cdf, exponential distribution, and normal distribution. So, in this lesson, we will cover the standard normal distribution, the t-distribution, and the chi-square distribution. So, in the last lesson, we saw the normal distribution with the representation of x. The representation of x n mu sigma square represented normal distribution.

$$z = (x - \mu) / \sigma$$

So now, in the standard normal distribution, we represent z of n 1 and 0. So here, the mean equals 1, sorry, 0, and the standard deviation equals 1. So, a normal continuous random variable with an expected value mean of 0 and a standard deviation of 1 is called a standard normal continuous random variable. The density curve associated with it is termed a standard normal curve.

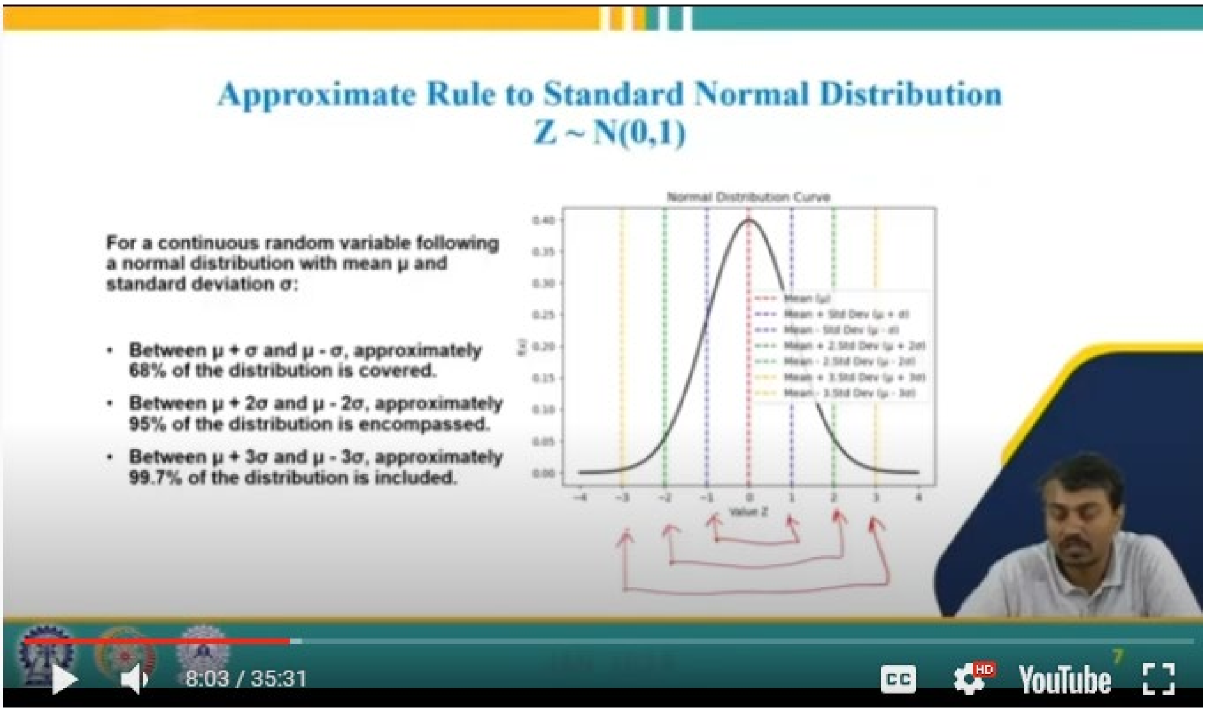


So now the mean is 0, and the curve is well spread between the, from the mean on both sides symmetrically. This is basically symbolized as a standard normal distribution curve with a mean equal to 0 and a standard deviation equal to 1. So there is a z, and a new variable is appearing; how do we calculate that z, and what is the z's relationship with the x variable, x continuous random variable? So let us discuss that. So this z is nothing, but z equals x minus mu and is divided by sigma.

So, if the population mean is known and the standard deviation is known for a particular value, we can estimate the z score. This technique is beneficial for evaluating a student's performance in different exams. For example, the student appears in the first--, second, and third-semester exams. Suppose the student score is known, and the standard deviation and the mean or the first, second, and third terms are known. In that case, whether the student is performing better in the first term, second term, and third term consecutively or progressing towards a better one, we can estimate using this z score. So, it is a beneficial technique, and there is an actual world application, particularly in academics. Also, it is helpful for other purposes. This standard normal distribution is valuable, and the z score is one of the essential features.

So, how do you calculate that? So, calculating the z score requires the population mean and the standard deviation. So, for example, here we have given the data x is equal to 57, and for that population, the mean is equal to 52, and the standard deviation is 4. So the z score is calculated as x minus mu divided by sigma, equal to 1.25. So, what does it indicate? It indicates that the selected value, which is 57, has a z score, indicating a 1.5 standard deviation from the mean. 1. standard deviation from the mean. So, this is the standard normal curve. So this is 0, 1, minus 1, and this is plus 1, minus 2, plus 2, plus 3, minus 3.

So, it is indicated that it is 1.25 standard deviation from the mean. It is here, around here. Approximation rule for the standard normal distribution Z of n 0 1, this is the standard representation. The standard deviation is 0, the mean is 1, the standard deviation is 1, and the mean equals 0.



So these are the representations. So, similarly, from 0 plus minus 1, this is the range. So, in the minus 1 to plus one range, this is 68 percent, similar to the normal distribution. Minus 2 to 2, this is 95 percent. Minus 3 to 3, it is 99.7 percent. That includes within the distribution. This is the same for the normal distribution. Also, you have seen sigma mu plus two sigma mu plus three sigma mu plus sigma like that. So, based on that, we can calculate the probability of Z from minus 1 to 1. Minus 1 to 1 is in 1 sigma. This area covers the one sigma coverage.

This is rightly pointed out, 0.68. For the two sigma minus 2 to 2, this covers 95 percent. Now, what is the probability of Z greater than minus 1? Z is more significant than minus 1, so this is mainly the area we are interested in, the total area. This is nothing but the area covered from this area to this total this side. What is this side? This is 68 percent, 0.68, and this side was 0.16, and this side was 0.16. So this is nothing, but 0.68 plus 1 is equal to 0.84.

So, what is the probability that Z is less than minus 3? This is nearly 99 percent; that is 1, that is less than minus three, is 0. Z minus 2, Z is equal to less than 2, so Z less than 2 is this value. So, this value signifies 95 percent, 0.95, and this side is 0.25, no, no, 0.025, 0.025, and this side is 0.025. So, less than 2, so this total area. So, this is 0.95 plus 0.25, this is 0.975.

So, let us discuss some of the properties of the standard normal variable, Z, the first property. So, what is the probability of Z less than minus X? So, we take the example of minus 2, the probability of Z less than minus 2. So, we are interested in this particular area. So, from the standard regular table, we can find the cumulative distribution value Cdf, which is the total value. This is the value. So, this value we can estimate.

So, value up to this, total value up to this, we can estimate. What are we interested in in this area? So, it can be represented as one minus the probability of Z less than X. So, we are now ignoring the sign. So, this is a total of 1 minus this value, indicating this particular value. So, this specific value and this particular value are the same.

So, this is the first property. So, P of Z minus X equals one minus P of Z less than X instead of the minus sign. So, this is another property we will discuss: probability between the range of A to B. So, A to B, if the B is more significant than X, both are positive, for example. So, for example, it is 2 to 3, 2 to 1.

So, we are interested in this particular range. So, we know these two values, and one is not. So, we can deduct that value to get this specific value. So, this is well represented: the probability of Z less than B minus the probability of Z less than A. So, this is a fundamental property.

Property third is Z, mod of Z, which means Z may be supposed to be minus 2. So, the mod is 2. So, for that, Z is greater than 2, greater than 2, and greater than suppose some value A. So, there is an essential formula. So, it has been found that for both cases, it is the summation of 1 minus P of Z minus A and one minus P of Z minus A.

So, it is two into one minus P of Z of less than A. So, for example, two is 2, Z, mod of Z greater than 2. So, we will estimate the probability of Z of A that is 2; it would be around 0.95, and 1 minus 0.95 is 0.05 into two that is 0.1. So, this is the value we are interested in that is more significant than this and more significant than this. So, this encompasses 0.1. Similarly, another property that is mod of Z less than A, mod of Z in the last slide we have seen, mod of Z greater than A, the fourth property is mod of Z less than A. So, this is interesting; this is nothing but two into the probability of Z less than A minus 1. So, this property will be beneficial during the probability calculations from the standard regular table because that standard typical table is available to us. So, this is the standard regular table and the score; this is a shortened version because a number of values are also available. This is on the Z point from 0 to 2.4 and in the two decimal places here, 0.00, 0.01, 0.02.

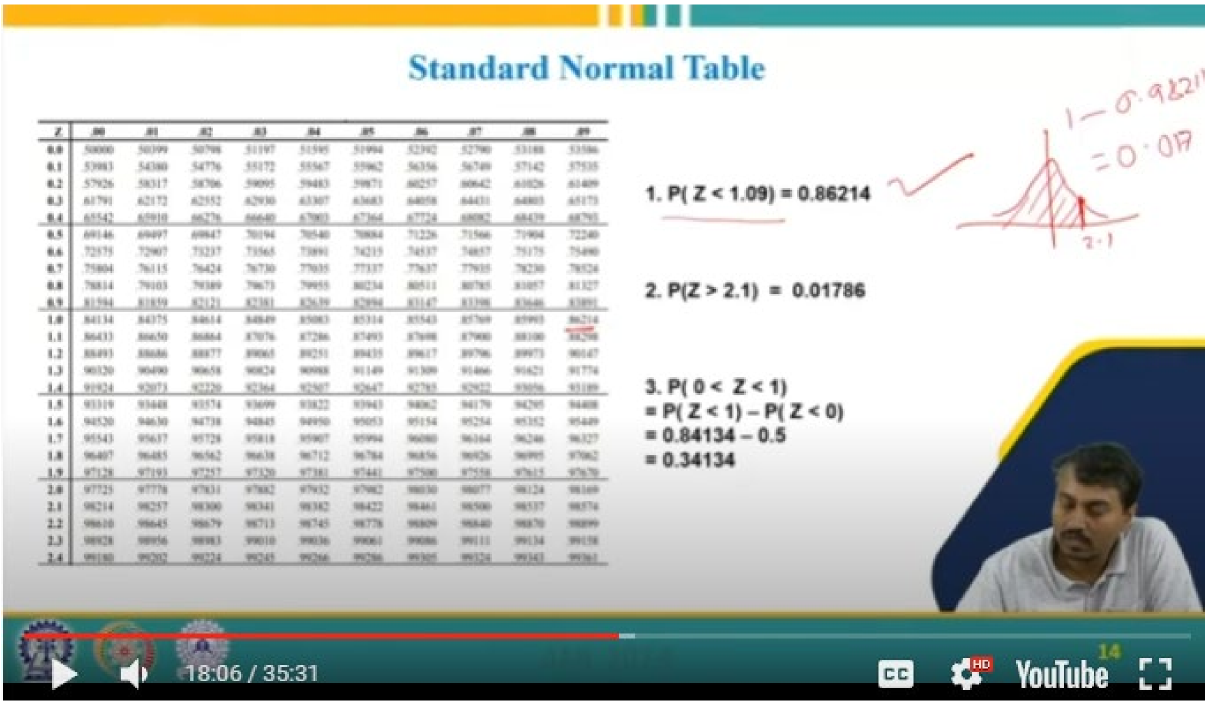
So, for example, if I want to estimate 0.11, if we want to estimate 1.96 1.96, this is a significant value, 97.5%, perfectly matching. So, 91.9, and the decimal is 0.06. So, from that, we are getting the standard average table value. So, let us estimate some of the values and how they are coming.

So, the probability of Z is less than 1.09, 1.09. So, this is 1, this is 0.09, 0.08214, 0.08214, perfect.

So, the probability of Z less than 1.09 is this. Now, what is the probability of Z greater than 2.1? More significant than 2.1 is this. So, this is nothing but from the curve if it is 2.

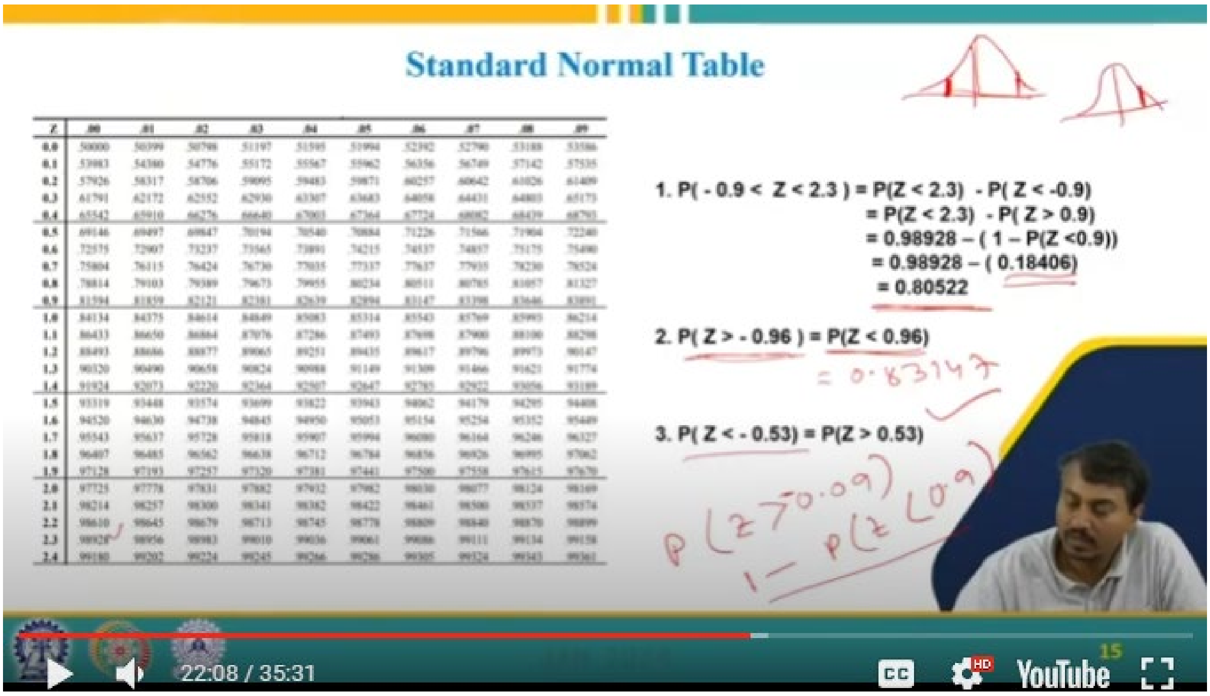
1, probability greater than 1. So, what will happen? What will we do? So, this particular value, 2.1, that we are getting from this particular data is this. So, we have to deduct that one minus that value.

So, 2.1 is this, 2.1 is this, 0.98214, 0.98214, nothing but 0.01786. Now, the probability of P is from 0 to 1. So, this is the curve, 0 to 1. So, this is 0, this is 1. What is the probability? It is around; if it is symmetric, the total was 68 persons, so it would be around 34 persons.



But from the table, it would be Z of, probability of Z less than one minus probability of Z less than 0. So, with a probability of 1, we can find that it is 0.84134, and the likelihood of Z0 is 0.5, precisely 0.34134. Another example is the probability of Z from minus 0.9 to 2.3, that is, Z greater than minus 0.9 less than 2.3, again the same concept. So, minus 0.9 is this particular value, and 2.3 is this range. So, the probability of Z greater than 0.09 minus, greater than this, is equal to the likelihood of more significant than this, which is equal to this total area.

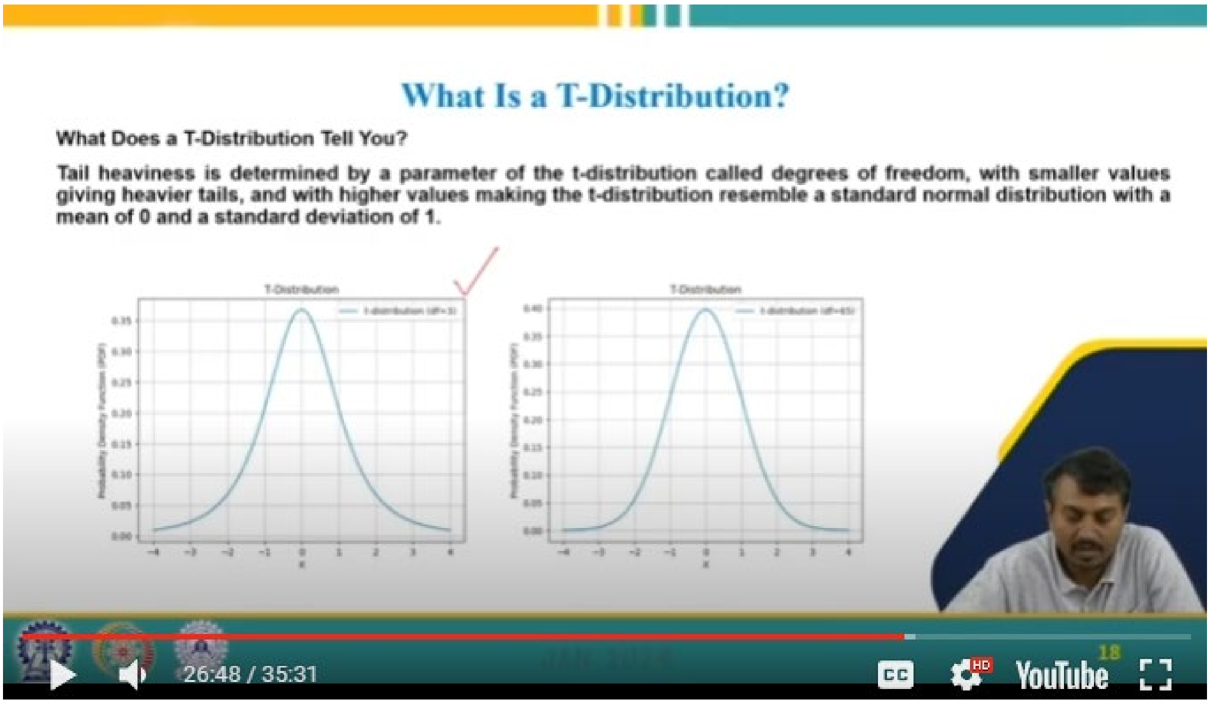
This can be found that one minus P of Z is less than equal to 0.9. That can be found out, so this particular value can be found out as 1.84, 0.18406, and the Z of 2.3, 2.3 is 92928, so we are getting this specific range of 0.80522. Now, what is the value of Z greater than minus 0.96? So, this is equal to Z of 0.96, similarly because this is the same value. So, 0.96 is around, is equal to 0.83147. So, for Z less than minus 0.53, it is a probability of Z greater than 0.53. So, a probability of Z greater than 0.53 is nothing but one minus the probability of Z less than 0.53. So, 0.53 is this, so one minus 0.70194.



So, this is the value of P of Z less than minus 0.53. Another is the mode that the property, the fourth property we can utilize, probability of the mode Z less than 0.2, is nothing but the likelihood of 2 multiplied Z less than 0.2. So, less than 0.2 is this; it is two into 0.57926 minus 1, this is 252, 8, minus one is equal to 0.15852. So, let us discuss another distribution, which is called t-distribution, and this t-distribution is often called student distribution. This distribution is similar to the normal distribution, which is bell-shaped. Still, it has heavier tails on both sides, an essential difference between the normal distribution and the t-distribution.

It also estimates the population parameter for small sample sizes or unknown variance. So, we use this distribution when the population variance or standard deviation is unknown. So, this is an important information to use the t-distribution. Second, the t-distribution has a greater chance for extreme values than normal distribution and, as a result, has flatter, fatter tails, and t-distribution is the basis for computing the t-test in statistics. We will elaborate on different tests of significance, and during that, we will use these t-test statistics.

The key takeaway from this t-distribution is that the t-distribution is a continuous probability distribution of the z-score when the estimated standard deviation is used in the denominator rather than the actual standard deviation because we do not know the standard deviation for the population. The t-distribution, like the normal distribution, is bell-shaped and symmetric. Still, it has heavier tails, which means that it tends to produce values that fall far from its mean, and t-tests are used in statistics to estimate the significance. So, what does t-distribution tell you? It tells heaviness is determined by a parameter of the t-distribution called degree of freedom, dF with smaller value giving heavier tails and with higher values making the t-distribution resemble a standard normal distribution with a mean of 0 and standard deviation of 1. So, here, in these two graphs, you can see the degree of freedom is three and the degree of freedom is 65.



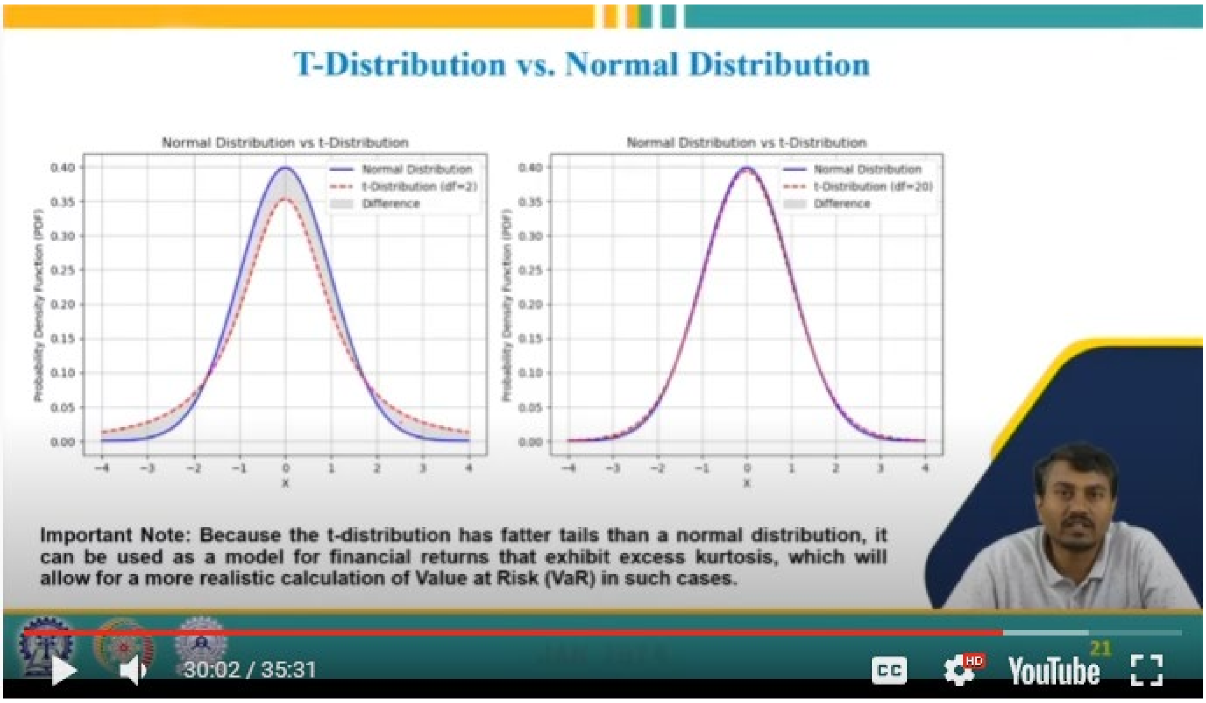
It looks ideally like the normal distribution. So, as the degree of freedom increases with the higher degree of freedom, it is more or less similar to the standard normal distribution. So, what is t-distribution? So, when a sample of n observation is taken from a normally distributed population having mean M, capital M, and the standard deviation d, the sample mean M and the standard deviation d will differ from M and d because of the randomness of the sample.

$$Z = (x - M)/D$$

$$T = (m - M)/\{d/sqrt(n)\}$$

So, a z-score can be calculated with the population standard deviation as of x minus M divided by D. The value z has a normal distribution with a mean of 0 and a standard deviation of 1. However, it uses the estimated standard deviation as a t-score calculated as t equals small m minus M divided by d of square root n. So, the difference between d, small d, and capital D makes the distribution a t-distribution with n minus 1 degrees of freedom rather than the normal distribution with a mean of 0 and standard deviation of 1.

So, let us compare these two distributions, t-distribution, and normal distribution. Normal distribution is used when the population distribution is assumed to be expected. The t-distribution is similar to a normal distribution with fatter tails; both take a normally distributed population. T-distribution thus has higher kurtosis than normal distributions. The probability of getting values far from the mean is more prominent with a t-distribution than a normal distribution.

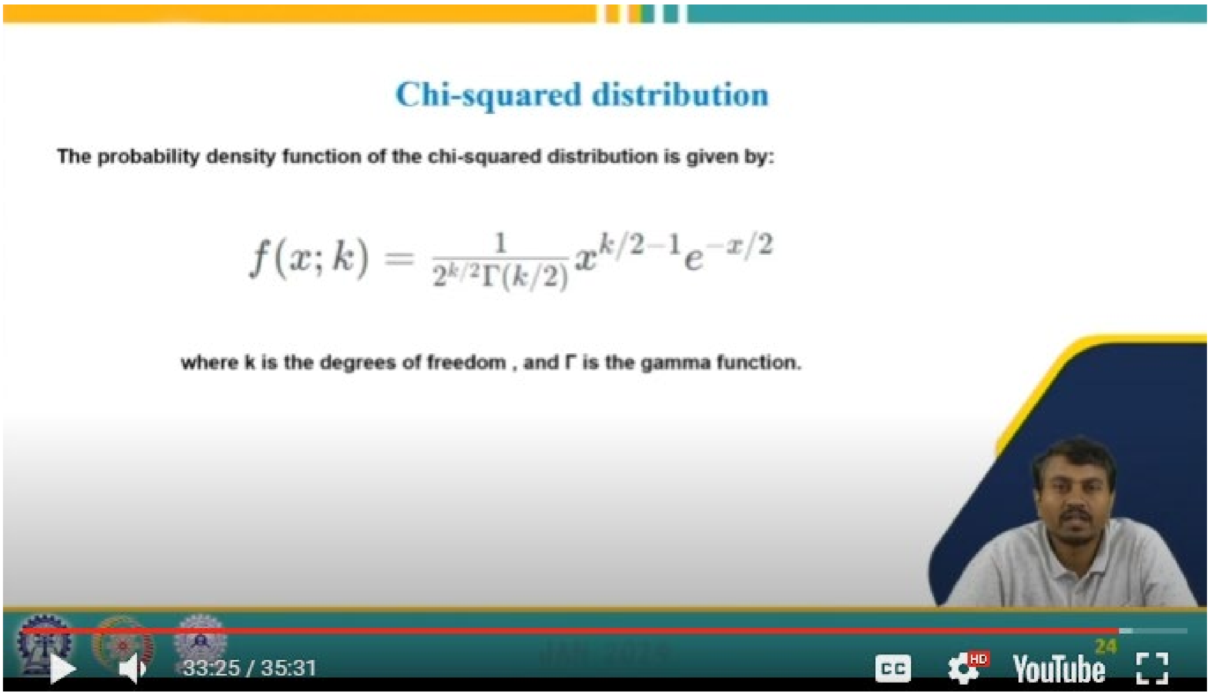


Here, you can see the normal distribution is the blue separate curve, and the dotted curve represents the t-distribution with a degree of freedom two, and the said zone defines the difference of area. This is so because the t-distribution has fatter tails than the normal distribution; it can be used as a model for financial returns that exhibit excess kurtosis, allowing for a more realistic calculation of value at risk in such cases. However, when the degree of freedom increases by 20, there is little difference between the t-distribution and the normal distribution. Limitations of using a t-distribution.

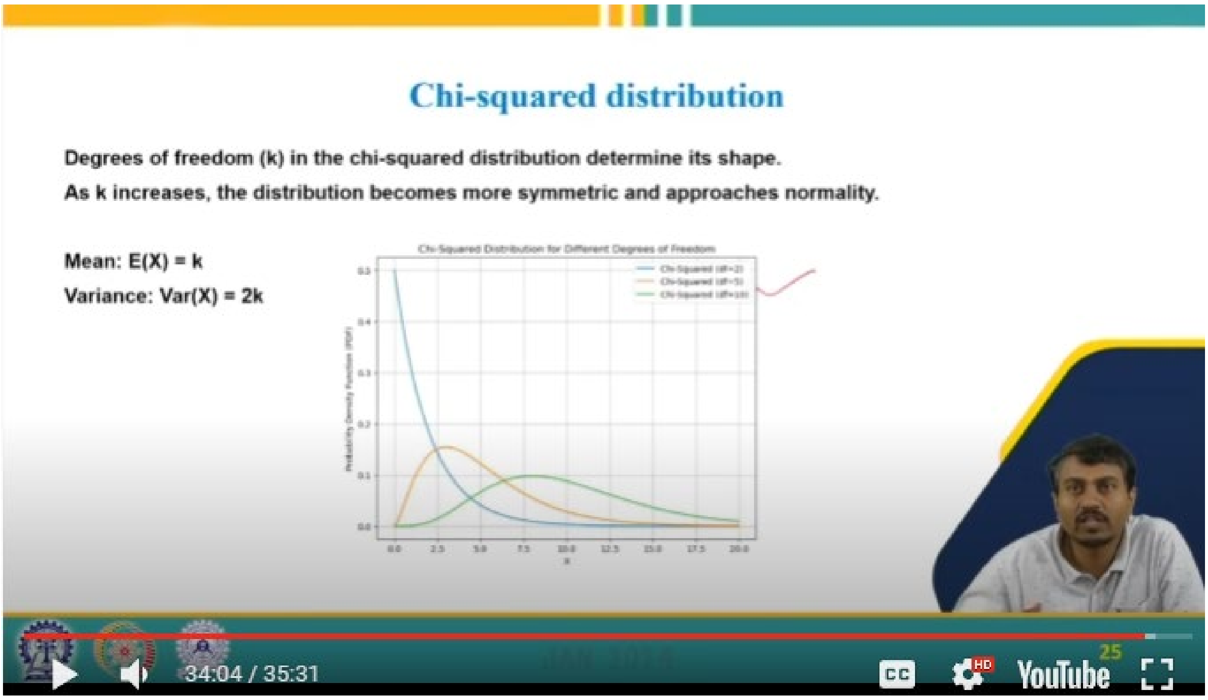
T-distribution can skew exactness relative to the normal distribution. Its shortcomings only arise when there is a need for perfect normality. The t-distribution should only be used when the population standard deviation is unknown. The normal distribution should be used for better results if the population standard deviation is known and the sample size is large enough. When should the t-distribution be used? The t-distribution should be used if the population sample size is small and the standard deviation is unknown.

If not, normal distribution should be used, which is the bottom line. T-distribution is used in statistics to estimate the significance of population parameters for small sample sizes or unknown variations. Like the normal distribution, it is well-separated and symmetric. Unlike normal distribution, it has heavier tails, resulting in a greater chance for extreme values. So, we will take care of the main application of t-distribution in the hypothesis testing lectures. So, we will elaborate on these t-testing methods in the significant test of significant test and hypothesis testing.

Chi-square distribution. The chi-square distribution is a continuous probability distribution in statistical inference, particularly in hypothesis testing and confidence interval construction. It is used in various statistical tests, such as the Chi-square test for independence and the Chi-square goodness of fit test. The chi-square distribution is a continuous probability distribution of the sum of squared standard average deviations represented by Chi-square. It is widely used in statistical hypothesis testing. The probability density function for Chi-square distribution is given by f of x k 1 by 2 to the power k by 2. Gamma function k divided by 2, x to the power k divided by two minus 1, e to the power minus x by two, and k is the degree of freedom.



So, the degree of freedom k in the chi-square distribution determines its shape. As k increases, the distribution becomes more symmetric and approaches normality. So, the mean expectation EX is k, and the variance is 2k. So, here, with the degree of freedom increasing, this is 10. So, for the 10, it is similar to the normal distribution. When the degree of freedom is increased, it lands in the standard distribution type.



Chi-square test for independence. One of the primary applications is the chi-square test for independence, which is used to determine if there is a significant association between two categorical variables, such as the chi-square goodness of fit test. Another application is the chi-square goodness of fit test, which tests whether an observed frequency distribution matches an expected distribution. The chi-square distribution is a particular case of gamma distribution about normal distribution. As k increases, the chi-square distribution approaches a normal distribution. We will take care of it. In the hypothesis testing lecture, We will discuss the chi-square distribution in detail.

So, these are the references. In this lecture, we have covered the standard normal distribution and its approximation rule. We have discussed the t-distribution and the chi-square distribution. Thank you.