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

Lecture - 18
The T-distribution

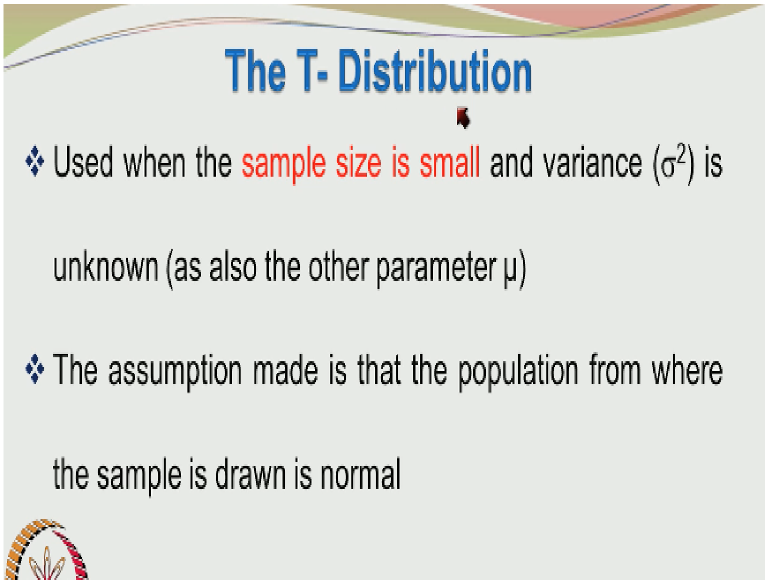
Hello welcome back. In today's lectures, we will be looking at 2 important distributions the first is the T distribution and the next is the Chi-square distribution. Both these distributions are extensively applied in design of experiments. Before this the most popular distribution we have studied is the normal distribution. The normal distribution had several desirable traits by now we know that it is symmetric mean median and mode coincide.

And it is obeyed by many sampling distributions of the mean another nice thing about the normal distribution is that the parameters of the distribution mu and sigma coincide with the mean and standard deviation. We also were looking at the sampling distribution of the mean and we were looking at a large sample size and we also were considering the situation where the population parameter sigma namely the standard deviation was not known.

So, hence it was recommended that we can substitute the sample statistic namely the standard deviation S instead of sigma. Being a large sample, the central limit theorem ensured that the distribution of the mean is normal approximately normal in fact. and we can substitute S for sigma. The natural question that arises is what will happen if I have a small sample size in many practical situations it may not be possible to have a large sample size.

A large sample size had many desirable properties it increased the precision while maintaining the level of confidence. It also enables non-normal populations to be considered in our analysis however we are now going to look at the case where the sample size is small, and the population variance sigma is not known.

(Refer Slide Time: 03:08)

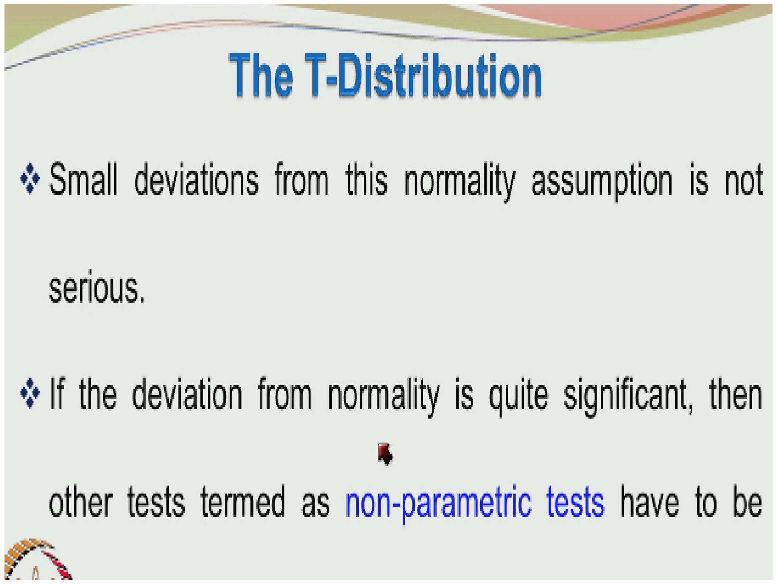


The T distribution sometimes also referred to as the students T distribution is applied when the sample size n is small. How small a small? is the query that immediately comes to us. Let us say that you have a small sample size of 6 10 or 15 only when the sample size exceeds 40 we can exploit the central limit theorem and also use S instead of sigma. But when the sample size is small then we have to use the T distribution.

So, let us know look at the properties of the T distribution the important assumption made is that the population from which the sample was drawn is a normal population. In other words, the distribution of the population members is normal okay this is a very important assumption however this is not a very binding or a very restrictive assumption in the sense that many of the populations behave close to normal type okay.

So, this is not very serious assumption.

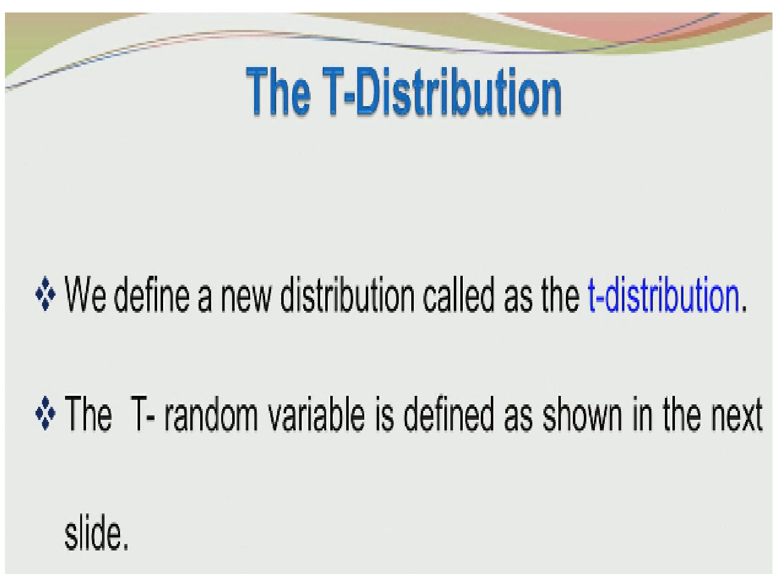
(Refer Slide Time: 04:55)



So, small deviations from the normality for the parent's probability distribution is permitted. However, what is to be done if the deviation from normality is quite significant then we have to go for non-parametric tests. These are quite interesting but beyond the scope of our current discussion the reason why we are doing the T distribution and the Chi-square distribution are as I said earlier they are extensively used in design of experiments methodologies.

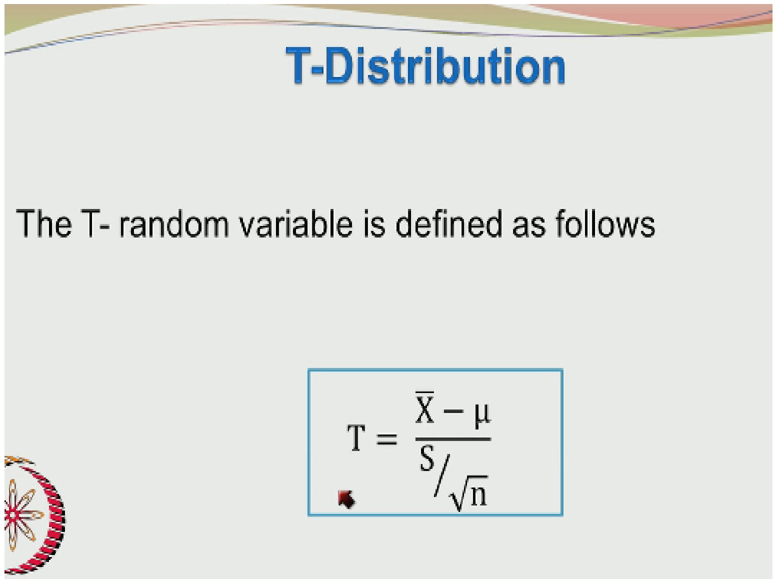
And also, when you look at any standard statistical analysis output you will find the confidence intervals noted the T values given the Chi-square distribution is the forerunner for the Fisher F distribution that we will see in the next lecture. Let us now focus on the T distribution.

(Refer Slide Time: 06:04)



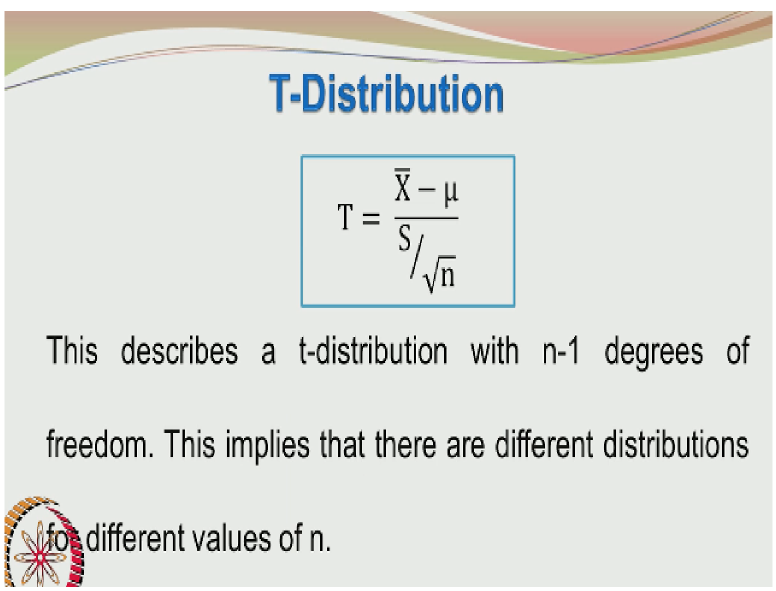
The T random variable is defined as shown in the next slide.

(Refer Slide Time: 06:16)



This is quite similar T is given by X bar-mu/S/root n well you may think that I have seen something like this before indeed we all have we have seen the standard normal variable being defined as X bar-mu/sigma/root n. So, that is an important difference Z was given in terms of X bar-mu/sigma /root n whereas T is defined in terms of X bar-mu/S/root n. So, this is enough to create the difference between the normal distribution and the T distribution.

(Refer Slide Time: 07:13)

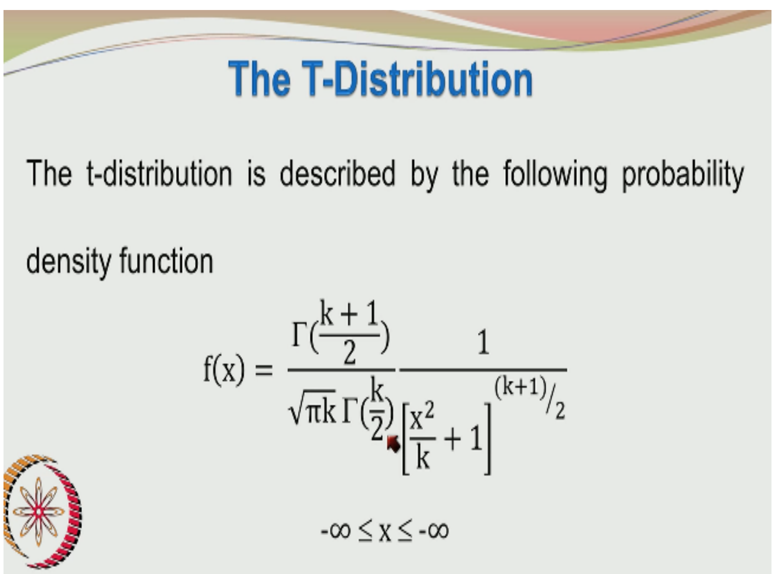


And the T distribution also is not a universal distribution we know that the normal distribution had parameters mu and sigma only. So, when you know mu and sigma that pretty much describe the normal distribution whereas when you are looking at the T distribution we have the parameter mu and then we also have another parameter called as the degrees of freedom. When you normalize the random variable X by subtracting the population parameter from it.

And dividing it by S/root n then the resulting random variable has a probability distribution which is centered at 0. This was the case also with the standard normal variable however the shape of the probability distribution for the T random variable depends upon the sample size okay. So, when you have different sample sizes the shape of the T distribution also changes hence the sample size is involved in the degrees of freedom.

In fact, we say that this is the T random variable which is following the T distribution with n-1 degrees of freedom again you might have seen this n-1 earlier. When you calculated the sample variance from the random sample data you divided the square of the deviations by n-1. Hence the same n-1 also figures as the degrees of freedom when describing the T distribution.

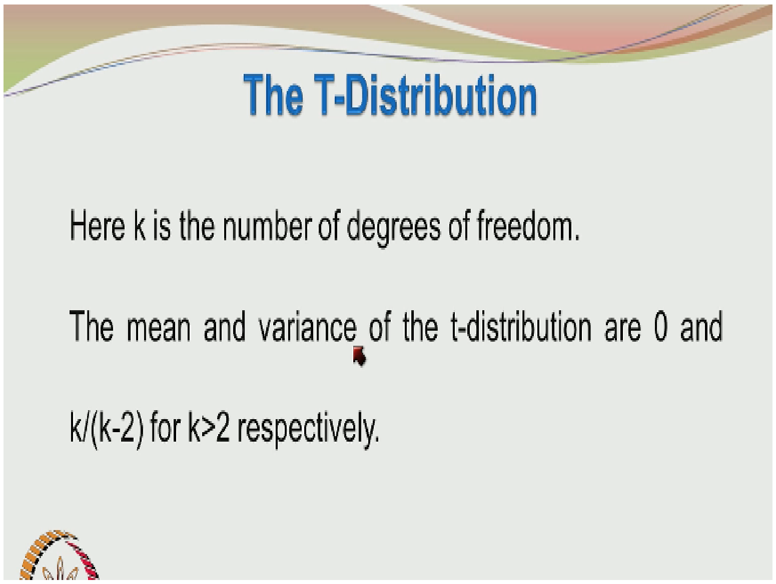
(Refer Slide Time: 09:57)



Well we will not be really using the probability distribution function for the T distribution it is beyond the scope of the present scope and objectives to play around with the probability density function f of x for completion I am just giving the form of the T distribution. It is gamma k+1/2 square root of pi k gamma k/2 x square/k+1 to the power of k+1/2 okay. So, what is this k? k is nothing but the degrees of freedom.

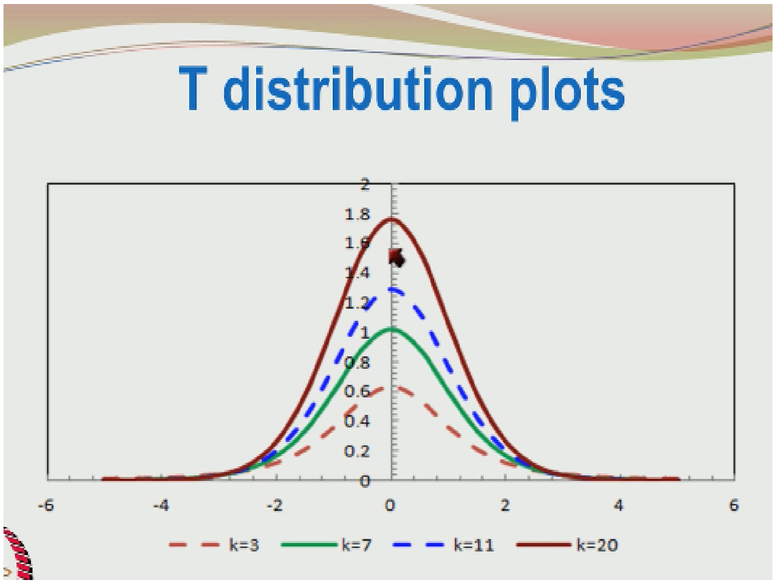
Since the T distribution is also having other applications we do not use n-1 in terming the degrees of freedom. So, we use a general parameter k but for our applications in statistics we can set k=n-1 but for general purposes the parameter k is the degrees of freedom associated with the T distribution.

(Refer Slide Time: 11:30)



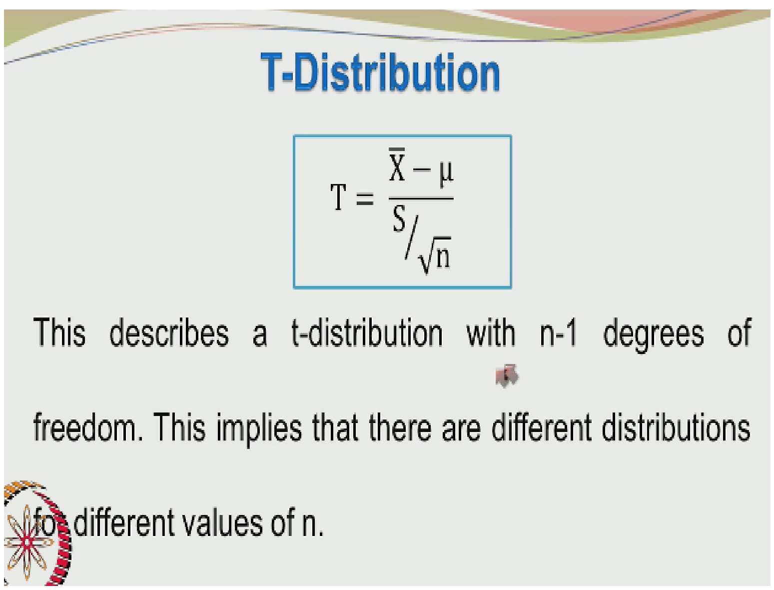
And another important thing this the mean and variance are not directly present in the T distribution as they were present as parameters in the normal distribution. In fact, the mean and variance of the T distribution are 0 and k/k-2 for k>2 respectively. What it means is the mean is 0 and the variances is k/k-2 of course k cannot be=2. So, the minimum degrees of freedom for the T distribution would be >2 it will be 3.

(Refer Slide Time: 12:13)



Let us look at the T distribution plots here they are you can see that several graphs have been drawn here and all these graphs are centered at the mean 0. The reason for that is we have centered the T distribution at the origin by making the transformation T=X bar-mu/ S/root n. So, this is what we have done when defining the T random variable.

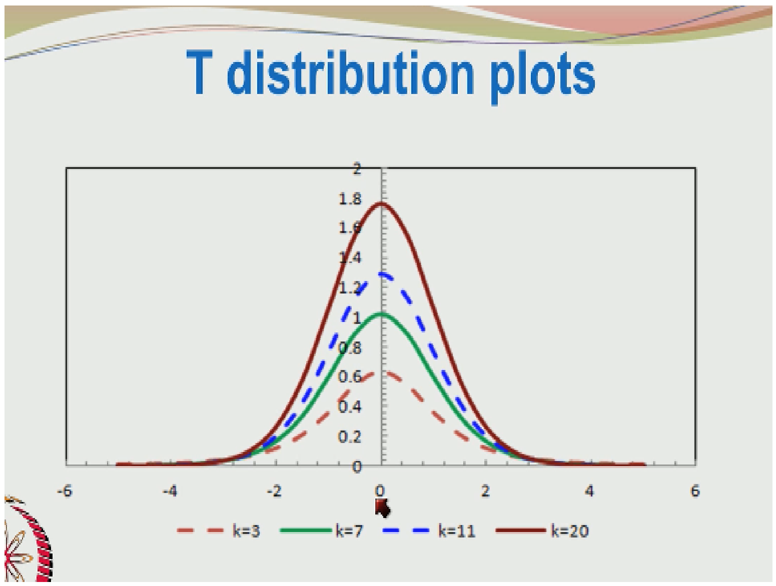
(Refer Slide Time: 12:56)



So, after you subtract by mu the distribution gets centered at the origin mu is the mean of the population. It is the average value of the population. Please note that we do not know the value of mu. So, the question naturally arises is in order to define the T random variable you have the sample mean X bar you have the sample standard deviation S and you have the sample size n, but you do not know mu then, how will you have a value for the T random variable?

So, many of our statistical analysis in the future will involve testing of the means. So, we speculate or hypothesize on the value of the population mean mu. So, we are given a value of mu which is assumed or postulated or hypothesized or speculated and that value we can substitute here. So, we will see more on this in the future classes right.

(Refer Slide Time: 14:39)



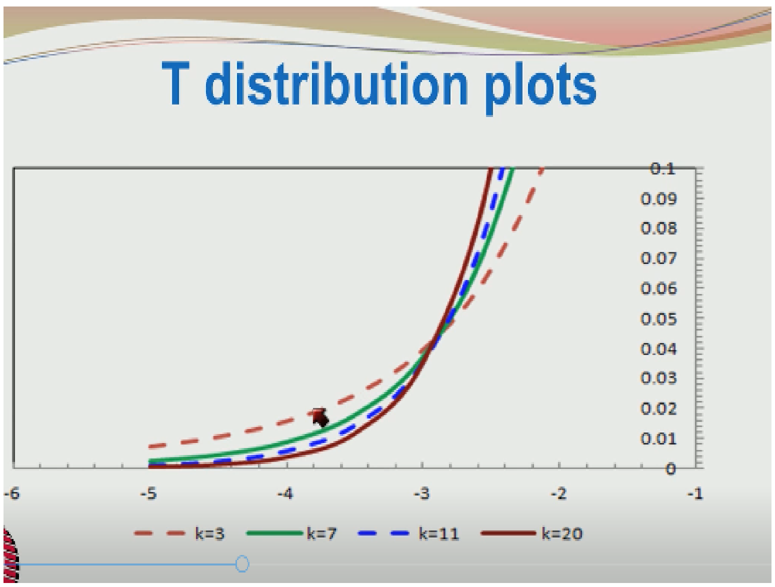
So, now we are having different curves whereas in the normal distribution we had only 1 curve and a person looking at this curve cursorily or superficially will say oh this looks like a normal distribution. In fact, the T distribution bears similarities with the normal distribution with the standard normal distribution. Because the mean is 0 and the distribution is symmetric 2 and-2 are equidistant from the origin.

So, if you go to 2 f of 2 will be=f of-2 so that is what is meant by the symmetric nature. The area enclosed by the curve is such that the area under the curve from 0 to let us say again 2 will be=to the area enclosed by the curve when X varies from 0 to-2. So, from 0 to-infinity the area under the curve is 0.5 from 0 to +infinity the area under the curve is again 0.5. So, that the total area under the curve=1.

That is the primary requirement of the probability distribution function. Now where did this curve come from these curves represent different values of k. One nice thing about the T distribution is it remains symmetric even for low values of k. There are a few other distributions which we will see presently that are no longer symmetric when the k value or the degrees of freedom decrease.

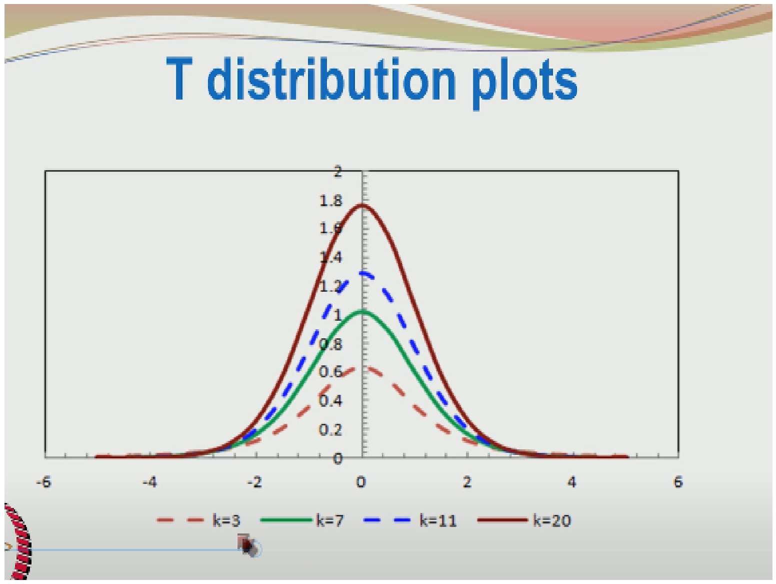
In this case for k=3 you have this brown curve and then for k=7 you have the green curve the blue curve is for k=11 and the dark brown curve for k=20. So, it can be seen that the curve becomes taller and the spread decreases when the degrees of freedom actually increases if you can see this portion I will highlight it you can see the brown slightly above the dark brown curve. So, you cannot see much from this figure.

(Refer Slide Time: 17:34)



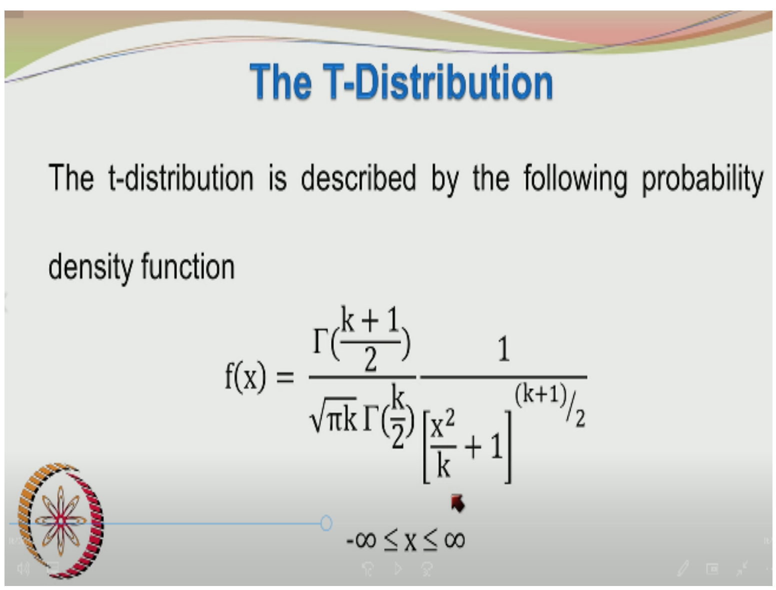
So, let us go to the next figure which will make things clearer, so the dashed brown curve corresponds to k=3. So, we are having a thicker tail for the T distribution and when the k value increased the curve became taller at the center, but it became thinner at the tail. So, this is for k=20 and this is for k=3. So, you can see that when the degrees of freedom is less then more probability is packed by the T distribution in the tail portion.

(Refer Slide Time: 18:33)



To summarize as the k value increased the distribution became tall and when the k value is small the T distribution became short at the center, but it became thicker in the tail regions. These curves can be easily generated by a spreadsheet okay what you all have to do is to use the probability distribution.

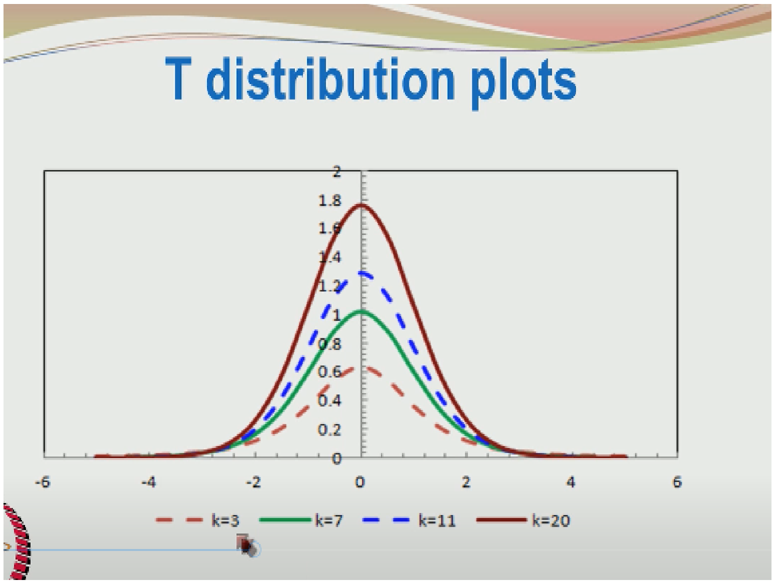
(Refer Slide Time: 19:07)



And you see the x which the independent variable is here you can take a very large value of x and a very small value of x. So, the range of x is from-infinity to+ infinity a small typo is there. I will just correct it. It should be from-infinity to +infinity and k is the parameter you can specify a value of k you can say k=3 4 whatever you want make sure k value is specified > 2 and then you can define this function in a spreadsheet.

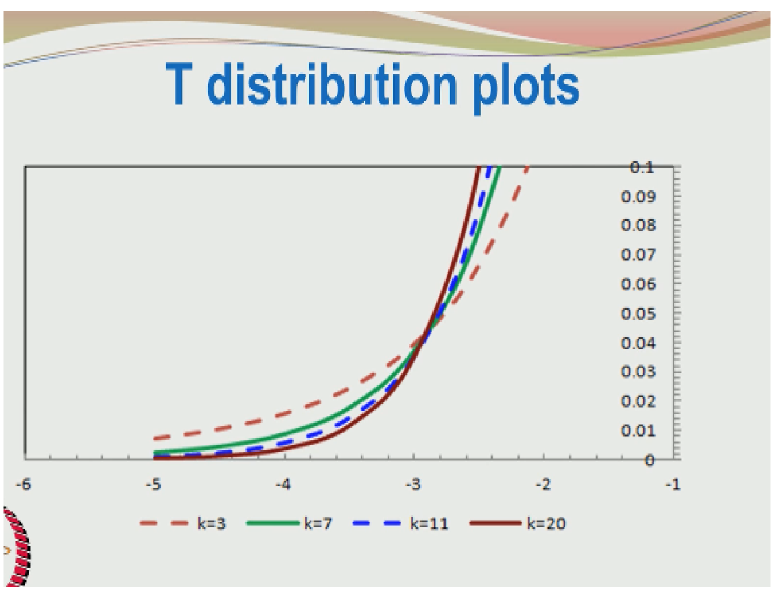
Vary the value of x from let us say a very small number-20 and then go up to+20.

(Refer Slide Time: 20:14)



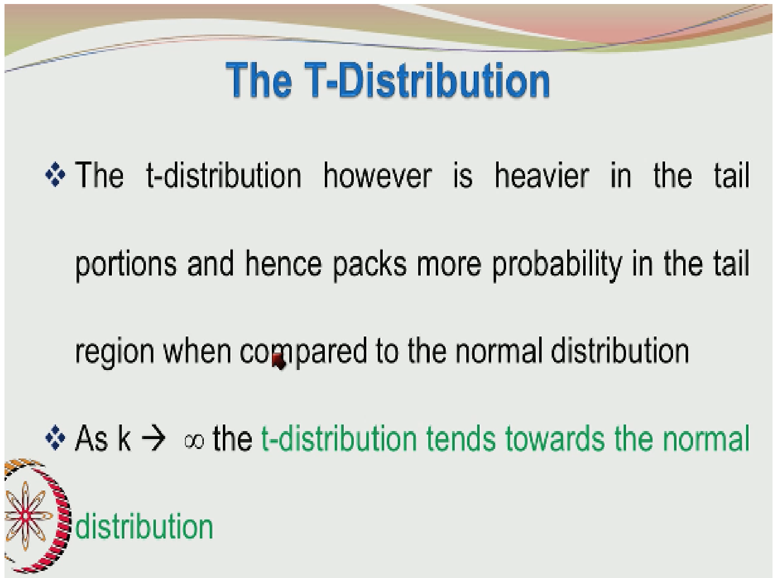
The values of the probability distribution function f of x reduce T fast you can see that even at k=3 if you go to-6 or+6 the value is pretty close to 0 and another thing you may notice is the curve changes its shape more and more slowly as the k value increases okay and beyond a certain value of k a very high value of k there would not be that much change in the shape of the curve.

(Refer Slide Time: 21:14)



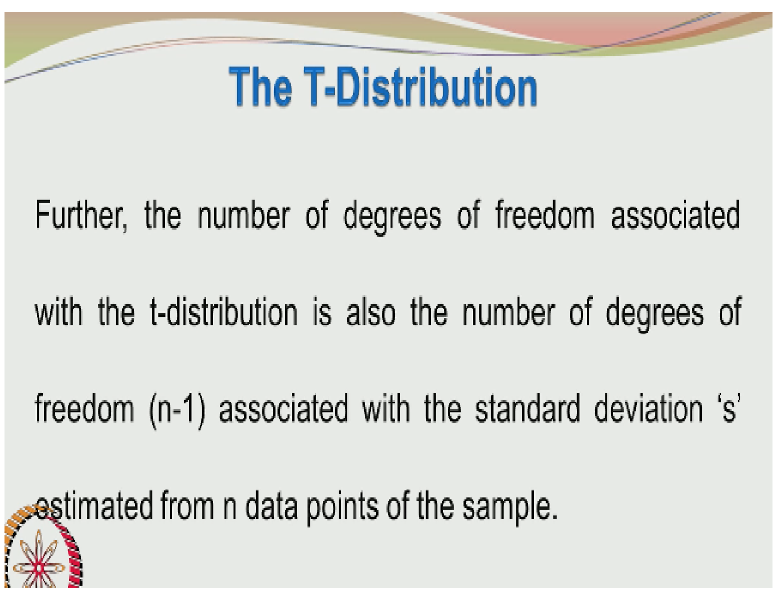
You can see that from k=3 to k=7 there is a big change whereas from k=11 to k=20 that change is not much. So, as I said earlier the T distribution resembles the normal distribution with 0 mean and both are unimodal symmetric about the origin and have a maximum value at the origin.

(Refer Slide Time: 21:48)



The t distribution is heavier in the tail portions and packs more probability in the tail region when compared to the normal distribution very interestingly as k tends to infinity that the distribution tends towards the normal distribution okay many distributions share this attribute. In the extreme the different distributions approach normality k tending to infinity is a mathematical criterion. But when k exceeds let us say 40 or 50 it pretty much is close to the values given by the normal distribution.

(Refer Slide Time: 22:33)

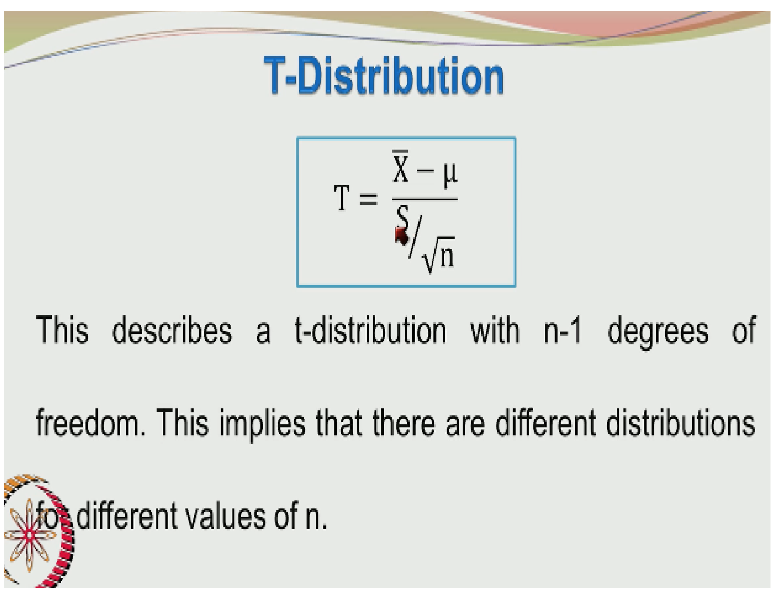


This degree of freedom when we use with the T distribution also matches with the degrees of freedom associated with the sample standard deviation S. The degrees of freedom mean with the T distribution also matches with the degrees of freedom with the sample standard deviation. Whenever we use the sampling distribution of the mean for small samples we have to use both the sample mean X bar and the sample standard deviation S.

And the sample standard deviation is associated with the n-1 degrees of freedom the reason for this we saw earlier in our discussions on the random variable and exploratory data analysis. We have n-1 degrees of freedom for the sample variance. The reason for that is not all the n deviations about the sample mean are independent only n-1 of them are independent. So, the sample standard deviation or the sample variance is based on n-1 degrees of freedom.

When we use the T distribution we also have k degrees of freedom when we use the T distribution for the sampling distribution of the means we use S in the definition of the T random variable.

(Refer Slide Time: 24:28)

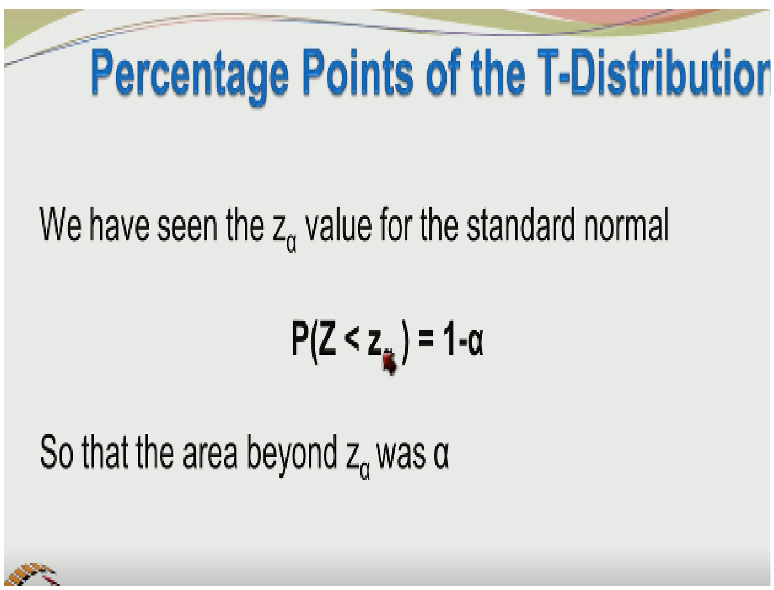


We use S in the definition of the T random variable, so k becomes=n-1 where n is the sample size. So, whenever we use the T distribution for the sampling distribution of the mean calculations for small sample sizes we use n-1 as the degrees of freedom for the calculation of probability values. Now where are these distributions used? In many practical situations we have to see the probability of a random variable following between 2 values.

Or being less than a particular value or being greater than a particular value. So, we have to find the probability. All the time we cannot use the normal distribution it depends on what distribution the random variable is following what distribution the random variable of our interest is following okay. Now when we talk about the distribution of the sample mean X bar for the case where the population parameter sigma is not known.

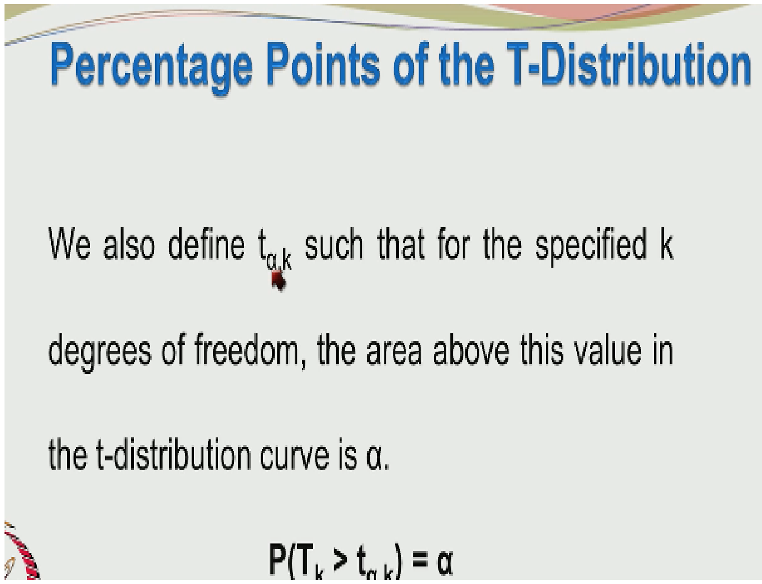
For the case where the sample size n is small and for the case where the parent population is following the normal distribution then we have to use the T probability distribution. So, whenever we are considering the probabilities of X bar laying between 2 values are $>$ than a particular value or $<$ than a particular value we have to use the T distribution. The probabilities have to be calculated using the T distribution. So, how do we get the probability values?

(Refer Slide Time: 26:47)



So, if z alpha is a calculated value of the standard normal variable we know that the probability of the standard normal random variable $<$ z alpha=1-alpha. So, that the area beyond z alpha was alpha. So, this is what we used in the case of a normal distribution.

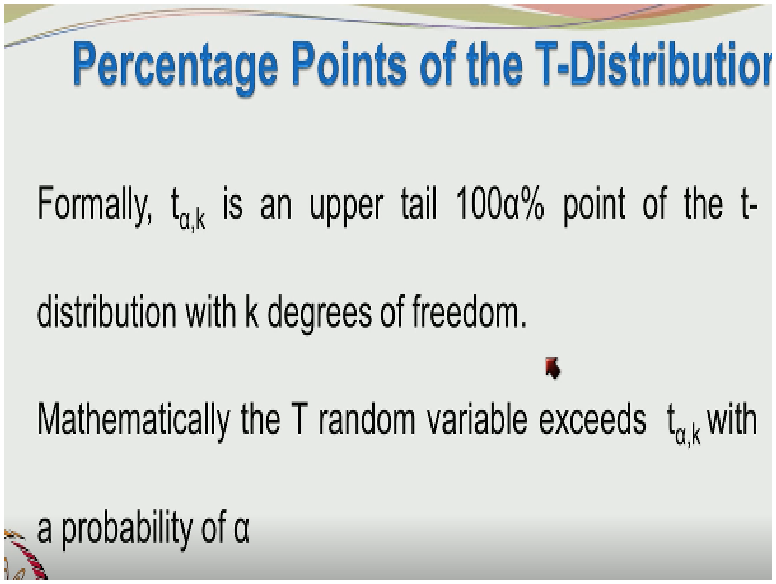
(Refer Slide Time: 27:24)



Now we define t alpha k such that for the specified k degrees of freedom the area above this value in the T distribution curve is alpha. Okay so now we are having t alpha and k, k is the degrees of freedom let alpha be a specified value. We have seen earlier that in the generation of the 95% confidence intervals the alpha value was given as.05 it may take values like.01 and.1 and so on but alpha=.05 was quite frequent or usual.

Now suppose we have the same alpha so what is the probability of the random variable T taking a value > t alpha k, what is the probability that the t random variable takes a value > t alpha k? and we define alpha in such a way that probability of T > t of alpha k=alpha. Okay so we define the subscript alpha in such a way that the T random variable exceeding t alpha k will take the probability of alpha. So, just note this definition we will be looking at the T distribution curve to understand this further.

(Refer Slide Time: 29:26)



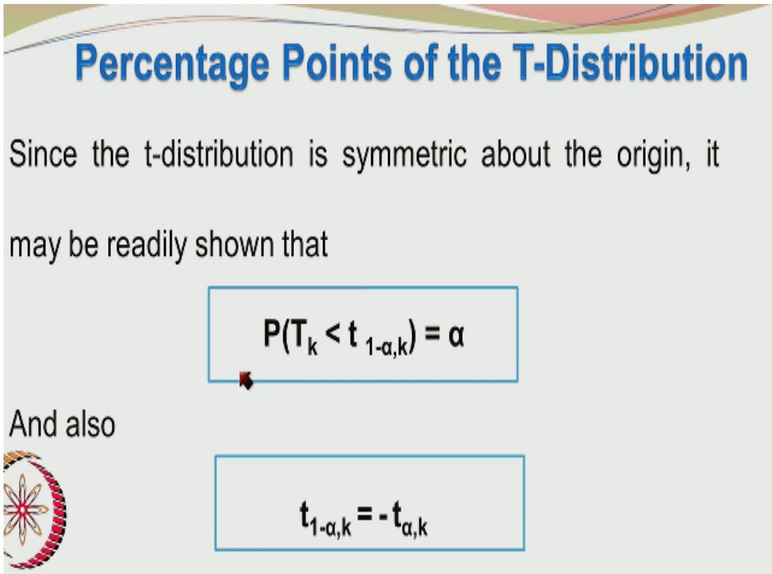
So, t alpha k is an upper tail 100 alpha% point of the t distribution with k degrees of freedom. This statistical terminology is fascinating and also important it is like grammar we have to use the correct terminology. So, again I repeat t alpha k which is used in the probability calculation it is usually a value okay it is represented in a general term t alpha k here it is a numerical value in fact.

So, t alpha k is an upper tail 100 alpha% point of the T distribution with K degrees of freedom what it means is whenever we are calculating the probability once the value of alpha is specified we look for the area under the curve beyond t alpha. Okay in the cumulative normal distribution probability chart involving the standard normal variable whenever a particular z value was specified the probability tables gave the area under the curve below the value of z alpha.

Okay but in the T distribution we are talking about the upper tail. So, whenever the value of alpha is specified we are looking at the probability in the tail region beyond the value of t alpha and the area under the curve beyond the value of t alpha is alpha itself. So, we can say that the area of the distribution or the area covered by the distribution below the value of t alpha would be=1-alpha the total area under the curve is 1.

The total area under the curve for any probability distribution by default by definition is=1. So, summarizing probability of the T random variable with a specified degree of freedom k taking on a value > t alpha, k=alpha. So, this is the definition.

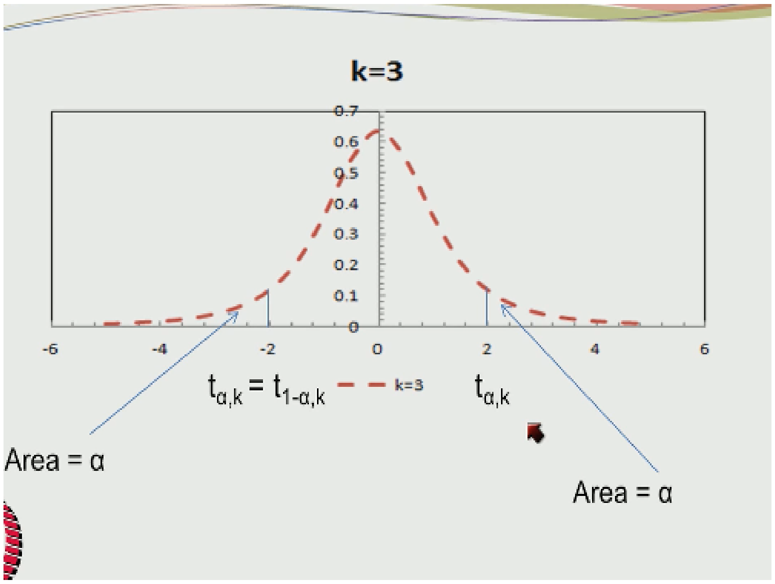
(Refer Slide Time: 32:12)



Now since the t distribution is symmetric about the origin it may be easily shown that probability of T<t1-alpha k=alpha as well. We are exploiting the symmetry of the distribution about the origin. So, if you locate a t 1-alpha k on the x axis of the T distribution the area below t1-alpha k he will be also=alpha. Now it may be also seen that t1-alpha k=k because the t distribution is centered around the origin.

So,1 1/2 of it corresponds to negative x axis values and the remaining 1/2 of it corresponds to positive x axis values. So, let us see how t1-alpha k=-t alpha k.

(Refer Slide Time: 33:30)



So, let us now look at this diagram this is the T distribution for a specified degree of freedom the specified degrees of freedom k=3 this curve again may be generated using a spreadsheet. Now let us fix the t alpha k which is a number at close to 2. Let me even make it at 2 well there is a slight discrepancy but let us assume that t alpha k=2. So, the upper tail probability will be alpha because this is some chosen value of alpha.

And so, the area covered by the upper tail is alpha. Now if you want the lower tail probability also to be alpha then you should locate t1-alpha k. The reason is the moment you put t1-alpha here the upper tail is going to be 1-alpha if the upper tail is 1-alpha then the lower tail value would be alpha. Okay what is the relation between t1-alpha k and t alpha k since t1-alpha k is located on this side on the left-hand side of the T distribution it would be a negative value.

But due to the symmetry of the T distribution t1-alpha k will be located at a certain distance from the origin and that distance will be same as the distance of t alpha k from the origin. So, we can say that t alpha k=-t1-alpha k okay, so this is the important relationship here. We can also say t 1-alpha k=-t alpha k okay, so this follows from the graph.

(Refer Slide Time: 36:25)

Percentage Points of the T-distribution for different Probabilities and Degrees of Freedom

	Probability									
DOF	0.40	0.25	0.10	0.05	0.025	0.01	0.005	0.0025	0.001	0.0005
1.00	0.325	1.000	3.078	6.314	12.706	31.821	63.657	127.32	318.30	636.61
2.00	0.289	0.816	1.886	2.920	4.303	6.965	9.925	14.089	22.327	31.599
3.00	0.277	0.765	1.638	2.353	3.182	4.541	5.841	7.453	10.215	12.924
4.00	0.271	0.741	1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.610
5.00	0.267	0.727	1.476	2.015	2.571	3.365	4.032	4.773	5.893	6.869
6.00	0.265	0.718	1.440	1.943	2.447	3.143	3.707	4.317	5.208	5.959
7.00	0.263	0.711	1.415	1.895	2.365	2.998	3.499	4.029	4.785	5.408
8.00	0.262	0.706	1.397	1.860	2.306	2.896	3.355	3.833	4.501	5.041
9.00	0.261	0.703	1.383	1.833	2.262	2.821	3.250	3.690	4.297	4.781
10.00	0.260	0.700	1.372	1.812	2.228	2.764	3.169	3.581	4.144	4.587
11.00	0.260	0.697	1.363	1.796	2.201	2.718	3.106	3.497	4.025	4.437
[illegible]	0.259	0.695	1.356	1.782	2.179	2.681	3.055	3.428	3.930	4.318
[illegible]	0.259	0.694	1.350	1.771	2.160	2.650	3.012	3.372	3.852	4.221
[illegible]	0.258	0.692	1.345	1.761	2.145	2.624	2.977	3.326	3.787	4.140
15.00	0.258	0.691	1.341	1.753	2.131	2.602	2.947	3.286	3.733	4.073

Now let us look at a typical probability chart for the T distribution probability table if you want to term it that way. So, the degrees of freedom are given here, and the probability value is specified here okay. It is slightly different from the standard normal distribution, so we have different probability values.4.25.01.05.025 so on to.0005. So, these probabilities are all upper tail values and the T value corresponding to this probability is given in these columns.

So, the T value with a distribution for 4 degrees of freedom packing an upper tail value of.05 is 2.353 for 3 degrees of freedom it is 2.353 and the T value for 4 degrees of freedom is 2.132. So, if my degrees of freedom for a given application is 4 and I want to find the T value that will have a probability in the upper tail as.05 then the required T value is 2.132. So. as the degrees of freedom changes the probability is fixed but the T values will change.

Okay so for a given probability as the degrees of freedom change the T values also change what is interesting to note here is for packing a specific probability in the tail portion okay the value of T decreases as the degrees of freedom increases okay. What it means is as the degrees of freedom increases the tail becomes thin and when the tail becomes thin you have to go closer to the origin to pack more area.

So, for a given probability the devalued decreases with increasing degrees of freedom. It is very interesting to relate the figure with the data given in the table by now you should know what is the z value corresponding to a probability of 0.05. In the cumulative distribution we talk about probability of.95 in the lower tail. In the T distribution we talk about the probability of.05 in the upper tail.

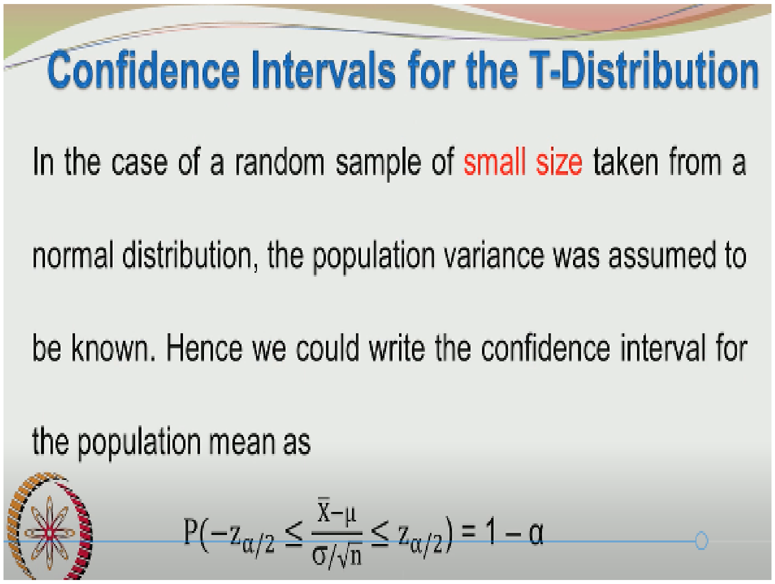
Let us see what happens to the T value when the degrees of freedom increases. So, we are now looking at.05 and it is coming to 1.697 at 30 and at 600 it is 1.660 it is at 40 it is 1.68 and that 600 it is 1.66. So, the change is not much and at infinity we are having 1.645,1.645 is a very popular number and so there is not much difference between the T value and the Z value because the Z value is the standard normal distribution variable value okay.

So, corresponding to a upper tail probability of.05 or a lower tail probability of 0.95 the Z value is 1.645 and the T value for a very large number of degrees of freedom is 1.660. So, we can see that the T distribution tends towards the normal distribution with increasing degrees of freedom another famous probability is 0.025. So, you are talking usually about the 95% confidence level and since we are talking about upper tail and the lower tail.

We have a probability of.025 in the right-hand side of the tail and 0.025 in the left-hand side of the tail. So, the z value corresponding to 0.975 cumulative probability in the lower tail or.025 in the upper tail would be 1.96. So, for the standard normal variable we have 1.96 for a T distribution with an upper tail probability of point 0.25 the degrees of freedom being 600 is 1.984.

So, for the T distribution with 600 degrees of freedom the upper tail probability of 0.025 will correspond to a T value of 1.984.

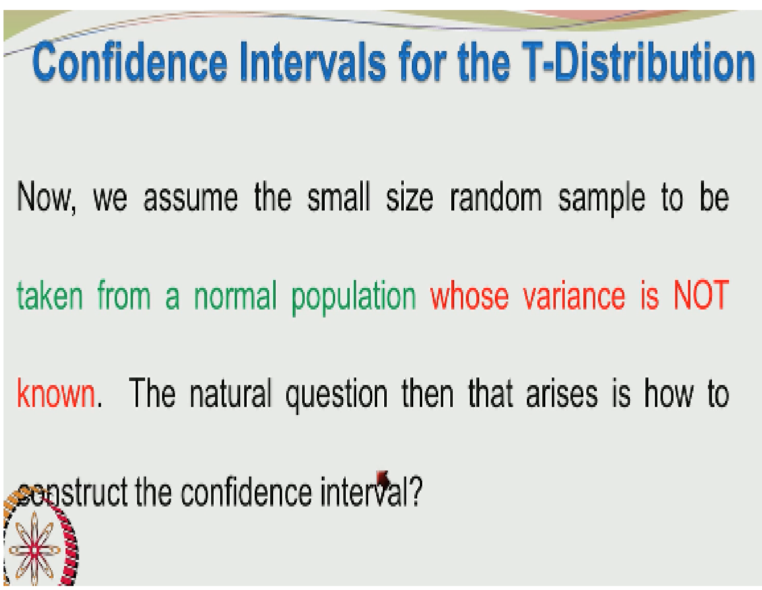
(Refer Slide Time: 43:17)



So, we can also generate confidence intervals for the T distribution we are taking a small sample size for the normal distribution we defined the confidence interval in this fashion. We were interested in obtaining the lower bound and the upper bound of the standardized random sample mean such that the probability value was 1-alpha. So, we were interested in finding out the lower bound-Z alpha/2 and the upper bound Z alpha/2.

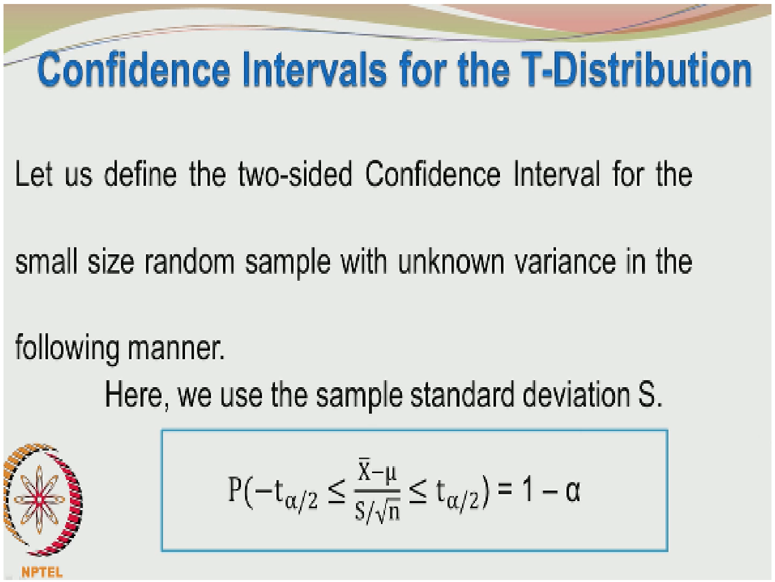
Such that the probability of-Z alpha/2 < X bar-mu/sigma/root n < Z alpha/2 was 1-alpha.

(Refer Slide Time: 44:16)



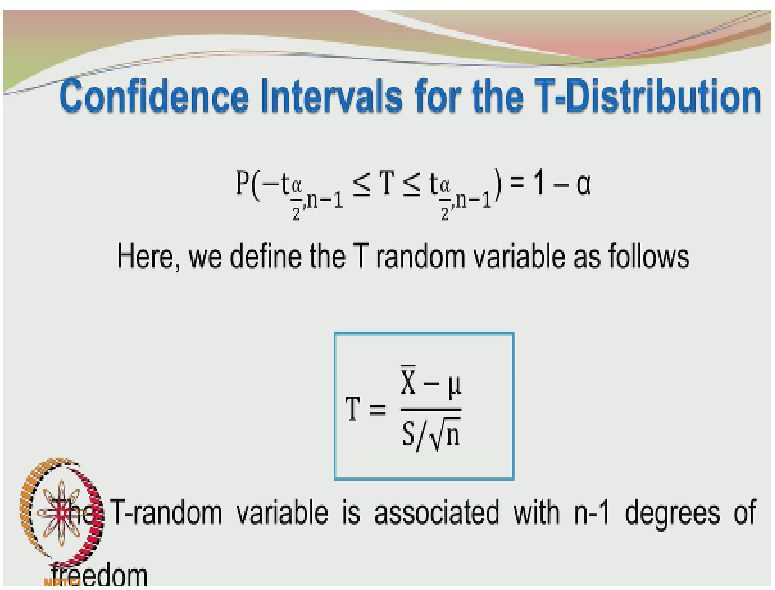
Now we are taking a random sample from a normal distribution, but the catch is the sample variance is not known. So, we are using the S value earlier what do we do for the construction of the confidence interval okay.

(Refer Slide Time: 44:40)



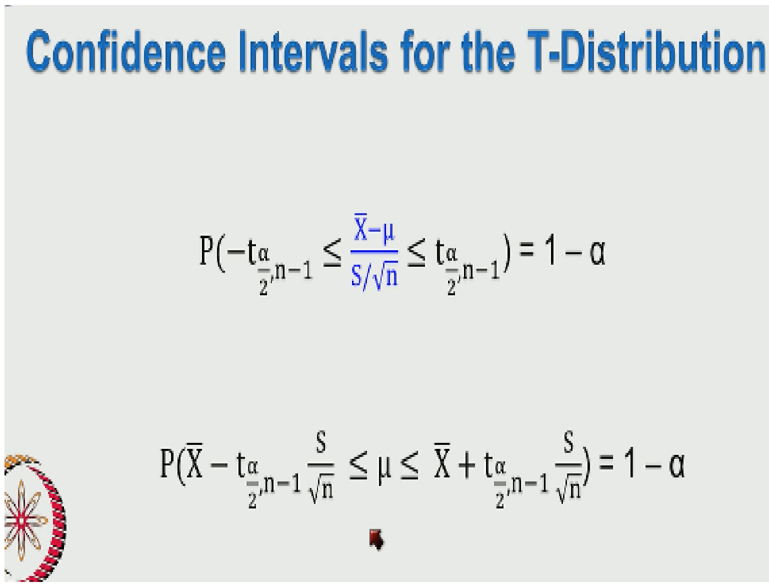
So, not surprisingly we use again the sample standard deviation S in constructing the confidence interval for the T distribution. So, we have probability of-2 alpha/2 <=X bar-mu/S/root n<=t alpha/2=1-alpha and the question is how I find-t alpha/2 and+t alpha/2 such that the probability of the standardized T random variable X bar-mu/S/root n lying between these 2 bounds will have a probability of 1-alpha.

(Refer Slide Time: 45:28)



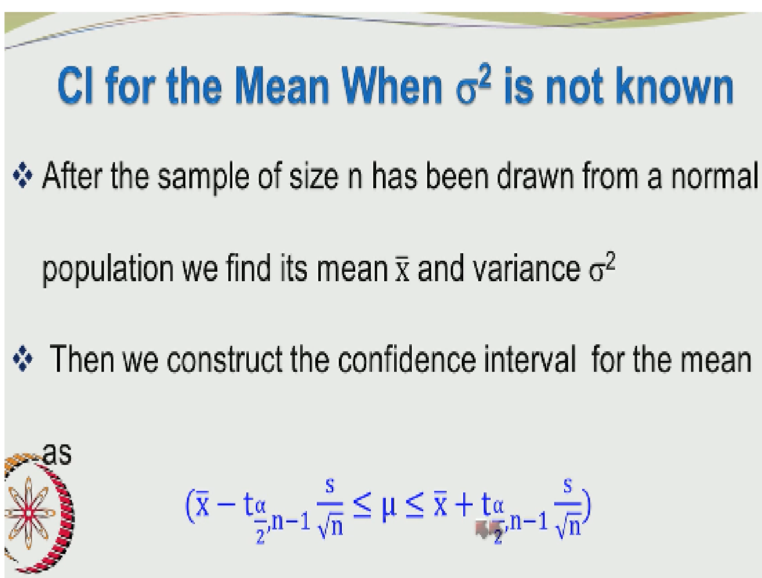
So, the T random variable we know is defined as X bar-mu/S/root n. So, we are having-t alpha/2 n-1 and t alpha/2 n-1 here and you also may recall that-t alpha/2 n-1 degrees of freedom is nothing but t 1-alpha /2 n-1. Because of the symmetric properties of the T distribution because of the symmetry properties of the T distribution.

(Refer Slide Time: 46:12)



So, this is the basic definition for constructing the confidence interval. We can manipulate on either side of the inequalities to eventually get a probability of X bar-t alpha/2 n-1 S/root n <=mu<=X bar+t alpha/2 n-1 S/root n. Please note the use of capital S and capital X in the representation of the sample standard deviation and the sample mean respectively and so this is what we have from this equation.

(Refer Slide Time: 47:00)

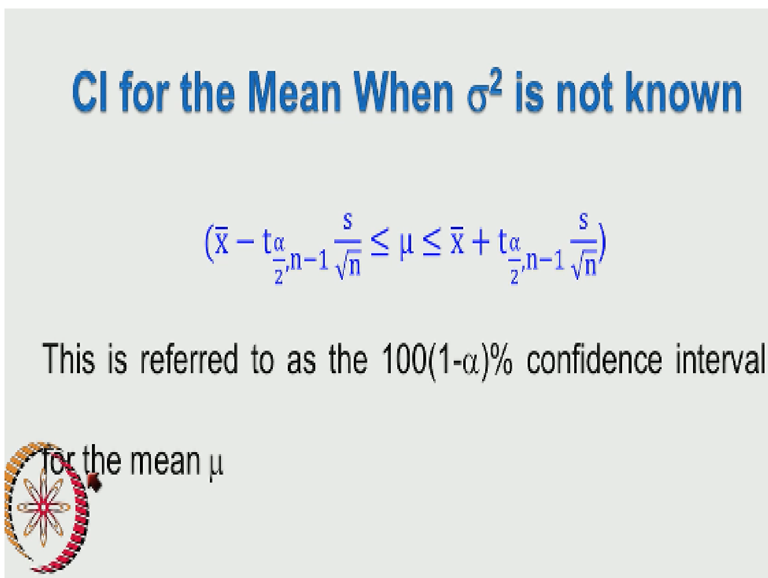


We can write the confidence interval for the population mean as X bar-t alpha/2 n-1*S/root n<=mu <=X bar+t alpha/2 n-1 S by root n. So, you are having a small sample the sample is taken from a normal distribution from an assumed normal distribution and then we not only want a point to estimate we also want an interval estimate and for constructing the interval estimate we define the T random variable.

The T random variable will in turn is defined in terms of the sample standard deviation S because the population standard deviation is not known. So, we have to define T as X bar-mu/S/root n and the degrees of freedom for the T distribution matches with the degrees of freedom used in the calculation of the sample standard deviation. So, once these are known we can easily construct the confidence interval.

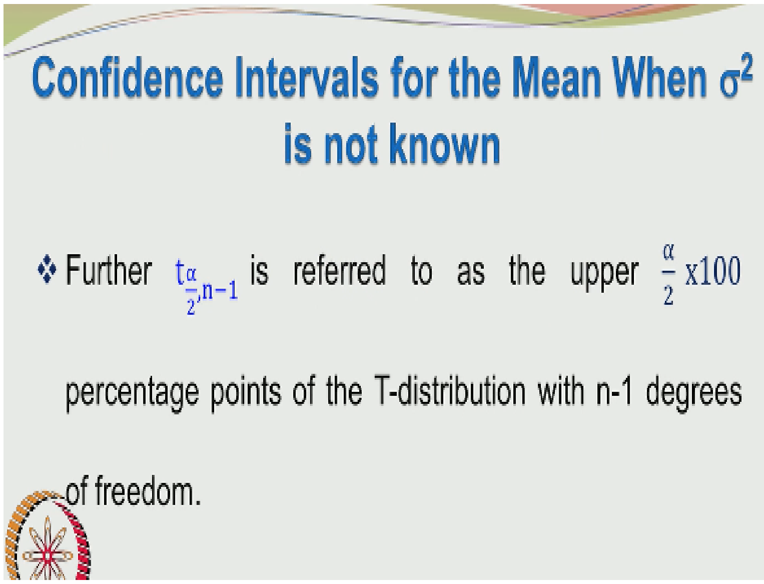
And the confidence interval is given by this expression.

(Refer Slide Time: 48:12)



So, this is referred to as the 100*1-alpha% confidence interval for the population mean mu. This is very important.

(Refer Slide Time: 48:28)



So, t alpha/2-1 is referred to as the upper alpha/2*100% points of the T distribution with n-1 degrees of freedom okay. There is a typo I will just correct it it should be % point not points it is a single point. So, this complete discussion on the T distribution we use it for small samples the parent population is normal, and the parent standard deviation sigma is not known parent population standard deviation sigma is not known. We will continue shortly.