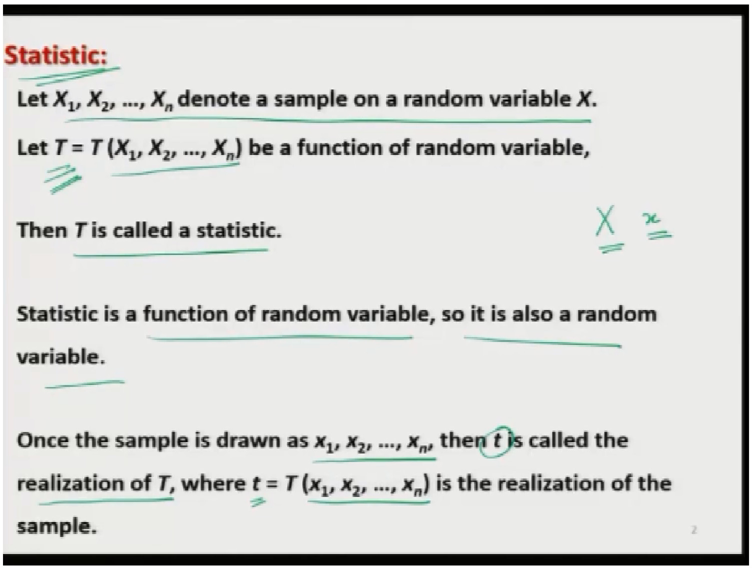
Essentials of Data Science with R Software- 1
Professor Shalabh
Department of Mathematics and Statistics
Indian Institute of Technology Kanpur
Lecture 50
Chi square Distribution

Hello friends, welcome to the course Essential of Data Science with R Software- 1, in which we are trying to understand the basic concepts of probability theory and statistical inference. So, from this lecture I am going to start a new topic this is Sampling Distributions. Well up to now, you have done probability density function, probability mass function for univariate, bivariate, multivariate random variables, these are also a sort of probability density functions, but they have got a different type of setup and different type of use.

Up to now as soon as I used to take any name then first I used to explain you the utility that under what type of circumstances they can be used but there are three sampling distribution which are called as chi square distribution, t distribution and F distribution these distribution are not used such as directly in a data set into an application, well they may be used but the type of things what we are going to do in the forth coming lecture, we are going to use them in taking out different types of statistical inferences.

So, we will use these distributions in some statistical methodology to take help, so that we can get the correct statistical outcome and its correct statistical interpretation. So, from that point of view these three sampling distributions are very important, they actually lay the basic foundation of the statistical inference, so now we are moving towards the statistical inference gradually, first I will try to take up some basic topics, fundamental topics which are essential to understand the methodologies for the statistical inference part. So, in this lecture let us try to understand what are the sampling distributions and in this lecture, I am going to talk about what is called the statistics and what is chi square distribution.

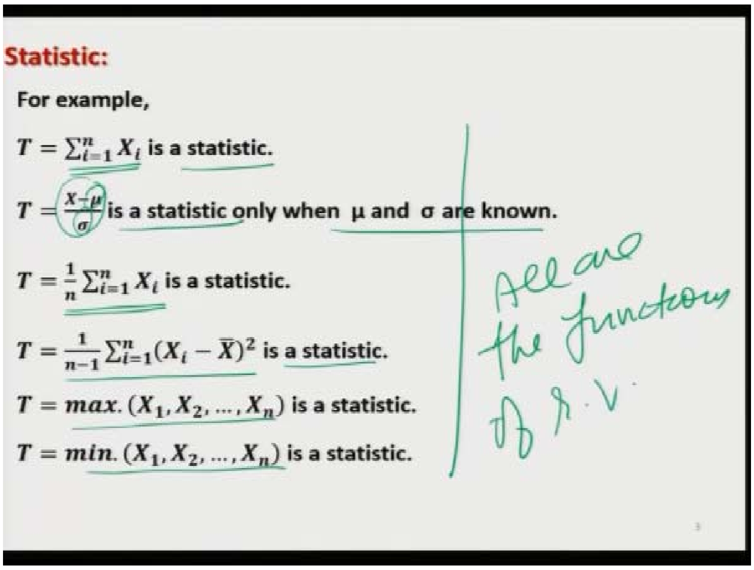
(Refer Slide Time: 2:33)



So, let us begin our lecture. So, now in this sampling distributions, first we have to understand a very fundamental definition this is a statistic, remember this is not statistics this statistic. So suppose, if I assume that $X_1, X_2, \ldots, X_n$ be a sample on a random variable X and then we try to create a function of these random variables. So, suppose T is the function of $X_1, X_2, \ldots, X_n$, then T is called a statistic and you can see here that this T, the statistic this is also a function of random variables, so it is also a random variable.

So, now we have understood two things if you remember X and x, X indicates the random variable, and the x indicates the value of the random variable. Exactly on the same way in case if I say that this is $x_1, x_2, \ldots, x_n$ this is a sample that is drawn on that variable X then t is called as a realization of T, where this t has got the same structure, same type of function but it is based on the sample values $x_1, x_2, \ldots, x_n$.

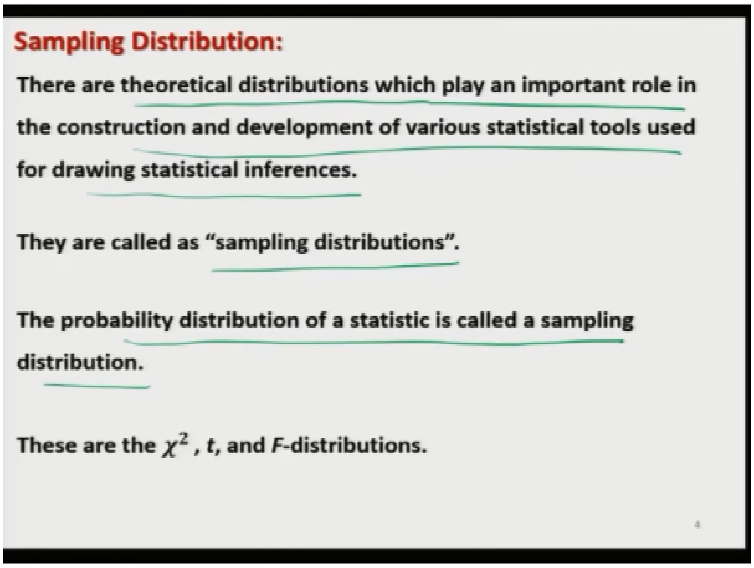
(Refer Slide Time: 4:18)



So now if you try to see there are many, many statistic that we already have used or we can actually use. For example, if I try to say $\sum_{i=1}^{n} X_i$ because the sum of all the random variable, this is also a statistic. In case if you try to take here the z value which you had indicated by X minus mean divided by standard deviation which is $\frac{X-\mu}{\sigma}$, this is also a statistic provided this μ and σ are known.

Similarly, if you try to take the sample mean $\frac{1}{n}\sum_{i=1}^{n} X_i$, this is also a statistic. Similarly, if you try to take this quantity sample variance $\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ this is also a statistic. Similarly, if you try to find out the maximum or the minimum values out of this X_1, X_2,…, X_n they are also statistics. So, you can see that all are the functions of random variable, so any function of the random variable is called as statistics, this is the one line definition which you always have to keep in mind.

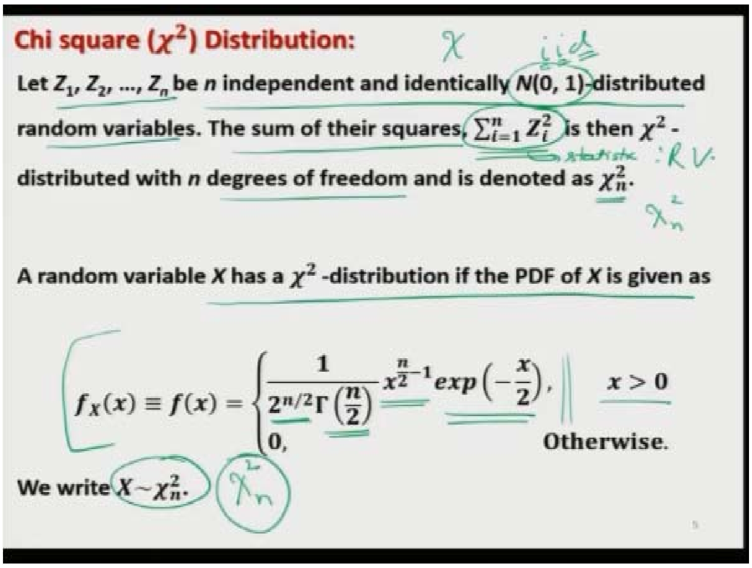
(Refer Slide Time: 5:33)



Now, I come to the aspect of sampling distributions, this sampling distribution are some theoretical distribution which play an important role in the construction and development of various statistical tools which are used for drawing the statistical inferences and they are called as in general sampling distributions.

And the probability distribution of a statistic is called a sampling distribution and when we are trying to define a particular type of sampling distribution, then the random variable corresponding to that distribution will also have to define in a particular way and that is what we have to understand, that in case if I am talking of chi square, t and F sampling distribution, so we have to first understand that what will be the random variable that is going to be a function of a random variable that is a statistic, so what statistics will follow the chi square, t or F distribution this is what I want to explain you.

(Refer Slide Time: 6:39)



So, first let me try to take here this chi square distribution. So, chi you know that is a Greek letter which is written here like this, so this is chi square, some time many people call it as say chi square, so this is not say chi or chi, this is chi. So, now first you have to understand that whenever I am going to talk about this chi square, t or F distribution, first I have to understand the statistic which is going to follow the chi square distribution.

So, suppose, if I say let $Z_1, Z_2, \ldots, Z_n$ be n independent and identically distributed $N(0, 1)$ distributed random variables. I have used here the symbol here Z because usually we have indicated Z to be a $N(0, 1)$ random variable. So, I try to take here n such independent and identically distributed random variables and this identically and independently distributed is also indicated as here iid.

So, i mean independent, i means identical, d means distributed, so as soon as we say that they are iid that means they are independent and they are identically, what is the meaning of independent that you understand, they are stochastically independent and what is the meaning of identical?

That means all these $Z_1, Z_2, \ldots, Z_n$ n they are coming from the same distribution $N(0, 1)$, it is possible that Z_1 is coming from $N(0, 1)$, Z_2 is coming from normal 1, 1, Z_3 is coming from normal 2, 3 and so on so in that case the distributions are going to be different for Z_1, Z_2, Z_3 and then we will say that they are not identically distributed, they may be independent that is a different thing.

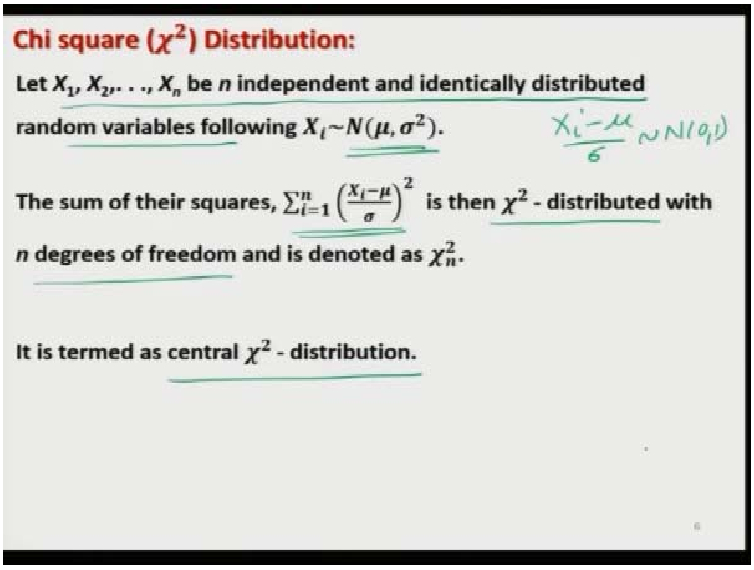
So, now if we try to consider such random variables which are $N(0, 1)$ then we try to find out the sum of their squares, $\sum_{i=1}^{n} Z_i^2$ so this is also a statistic that you can say and hence this is also a random variable, then we say that this quantity $\sum_{i=1}^{n} Z_i^2$ follows a chi square distribution with n degrees of freedom, and it is indicated here like this chi square and then here in the subscript this is here n, so n is going to indicate the degrees of freedom.

Now, the next question comes what is the degrees of freedom? So degrees of freedom in sampling distribution they try to control the characteristic, they try to control their behavior, their graphics and the concept of degrees of freedom in statistics can be defined in different ways but, yes, that area is possibly out of the view of this course but as soon as I try to take some example the meaning and the use of degrees of freedom will become clear to you, so you have to just wait for couple of minutes.

So, now in general I can say that a random variable x has got a chi square distribution if the PDF of X is $f(x) = \begin{cases} \frac{1}{2^{n/2}\Gamma\left(\frac{n}{2}\right)} x^{\frac{n}{2}-1} exp\left(-\frac{x}{2}\right), & x > 0 \\ 0, & \text{Otherwise.} \end{cases}$. And we indicate it by writing here $X \sim \chi_n^2$.

One thing I can assure you before moving forward that if you try to look into the forms of the mathematical functions of chi square, t or F distribution they may look little bit complicated but you do not have to worry for them because we are not going to use these forms anywhere and you need not to always remember them and you will see in the usage that we will simply be using one information that a random variable is following a chi square distribution with degrees of freedom and like this one only and this I will try to illustrate you and I can assure you quite in the forthcoming lectures at no place we are going to use it directly.

(Refer Slide Time: 11:04)

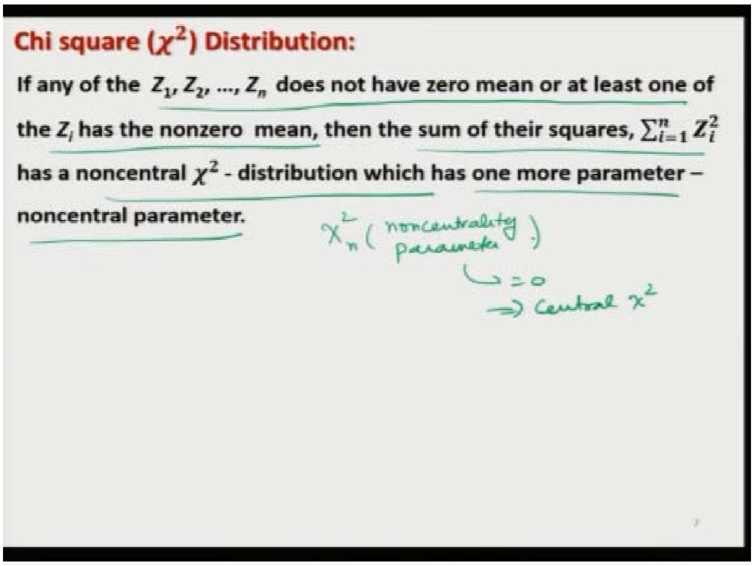


So, now if you try to see here in this result where I am assuming that $Z_1, Z_2, \ldots, Z_n$ they are an iid N(0, 1) distributed random variable, there I am assuming that the mean is 0 and variance is 1 but suppose if I take these distributions to be different. For example, if I say that let $X_1, X_2, \ldots, X_n$ be n identically and independently distributed random variables following X_i to be (μ, σ^2), where the mean is not equal to 0 and , $\sigma^2$is not equal to 1 necessarily. Then in that case what we try to do we know that if I try to take here Xi minus mu upon sigma then this will always follow a N(0, 1) distribution.

So, what we try to say in cluster case that $\sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2$ follows a chi square distribution with n degrees of freedom. And this is actually term whenever we are trying to take the such N(0, 1) random variable the chi square distribution is termed as central chi square distribution.

So when there is a central distribution, that means there should also be something called non central chi square distribution, but one thing I would like to make it clear that in most of the cases we are going to use only the central chi square distribution and in the common language of statistics whenever we are saying that the distribution is chi square, so unless and until we are writing or saying that it is non central chi square we assume that the distribution is central chi square.

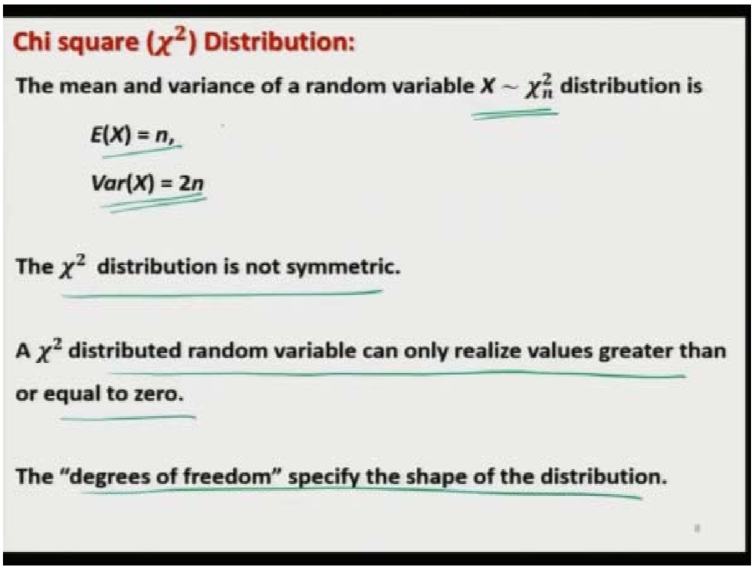
(Refer Slide Time: 12:51)

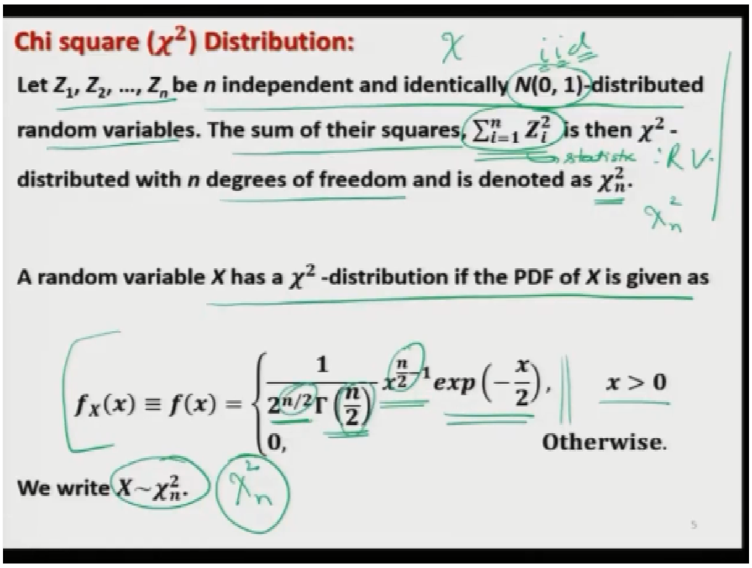


So let me try to give you an idea that what is this non central chi square. So, if any of the $Z_1, Z_2, \ldots, Z_n$ does not have zero mean or at least one of the Z_i has the non zero mean then the sum of their square that is $\sum_{i=1}^{n} Z_i^2$ has a non central chi square distribution and when we are talking of non central chi square then this distribution is going to be characterized by one more parameter which is called as non central parameter or non centrality parameter and we try to indicate it like this here a chi square n and inside the parenthesis we can write down the value of non centrality parameter.

And if a non centrality parameter is equal to 0 that means we have a central chi square, as simple as that, although we are not going to handle here the non central chi square, t or F but it is important for you to know what is this thing so that if needed you can use it, you can read from the book and you can use it and implementation of this non central in the R is very simple actually.

(Refer Slide Time: 14:07)



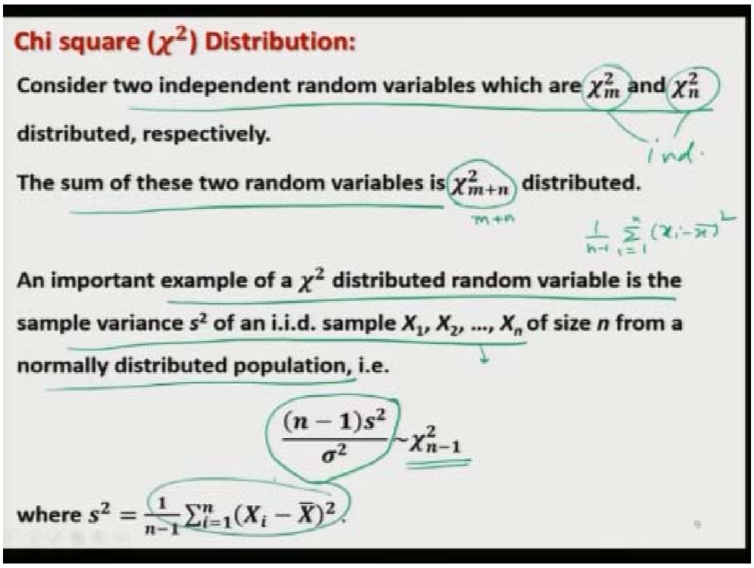


Now, some properties of this chi square distributed random variable, in case if I have a random variable X which is following a central chi square distribution with n degrees of freedom its mean is given by n and variance is given by twice of n, and this chi square distribution is not symmetric and the values of chi square are going to be greater than 0.

So, a chi square distributed random variable can only realize values which are greater than or equal to 0 and this degrees of freedom they specify the shape of the distribution and you can see from the probability density function that there is here is an n by 2, n by

2, n by 2, so there is an involvement of your n and this summation is also going up to n, this mean and variance they are depending on n and similarly if you try to plot their curve they will be changing with respect to the value of n.

(Refer Slide Time: 15:05)

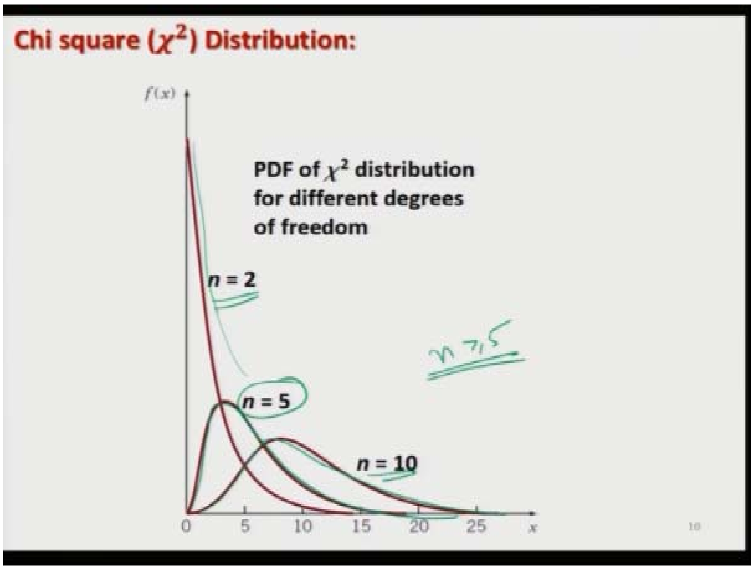


Now, an important property that is an additive property of chi square random variables, that if we consider two independent random variables which are distributed as chi square with m degrees of freedom and chi square with n degrees of freedom, so both these are independent, then the sum of these two random variables will also have a chi square distribution with m plus n degrees of freedom.

That is the first result I would like to inform you this is very important we are going to use it at different places and second a very important result is that the sample variance, do you remember that you had found the sample variance by this quantity $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$, this is related to chi square distribution, how? So, an important example of a chi square distributed random variable is the sample variance s square of $X_1, X_2, \ldots, X_n$ which are iid identically and independently distributed.

So, this $X_1, X_2, \ldots, X_n$ they are iid they are obtained from normal distribution and then we can say that $\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}$ distribution with (n – 1) degrees of freedom, where your sample variance s square is given by this quantity. So, that is also a very important result which is going to be used at many places in the estimation of parameter, test of hypothesis and different properties of statistics.

(Refer Slide Time: 16:55)

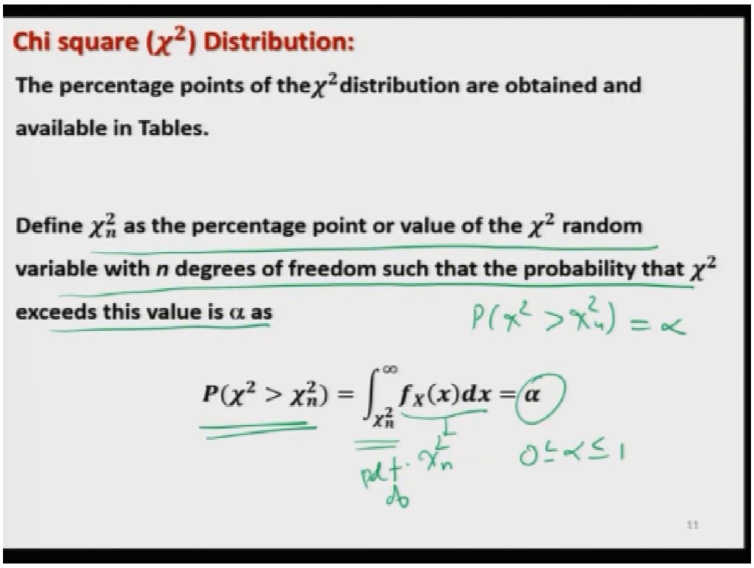


Now, if you try to see, just for your information, if you try to plot the probability density function for different degrees of freedom that will look like this. For example, you can see here for n equal to 2 this is here like this one, shaped curve and if you try to increase the degrees of freedom up to 3, 4 etc. this will more or less be like this structure but as soon as you go to n equal to 5 the shape of the curve changes it becomes like a closed curve.

And if you try to increase the n after this for example at n equal to 10 the curve will still be like as a closed curve, so that is a very important result what you have to keep in mind because that is going to be extremely useful in the statistical inference and usually you will see that we try to make an assumption that n is greater than or equal to 5, then my result and if n is smaller than 5 then we have to do some special things whenever we are trying to draw the statistical inference.

So, at that moment when I will use it I will inform you but I will refer to the same result that from this curve you can observe that as soon as n becomes greater than or equal to 5 the shape of the curve changes.

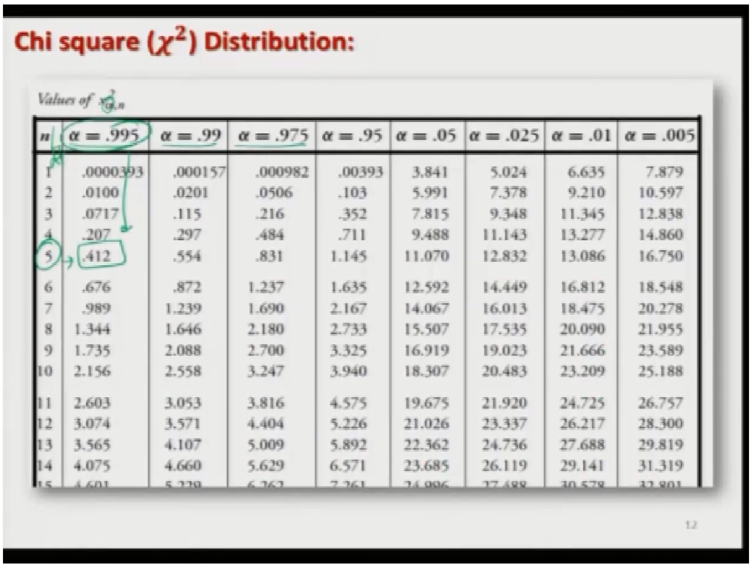
(Refer Slide Time: 18:16)



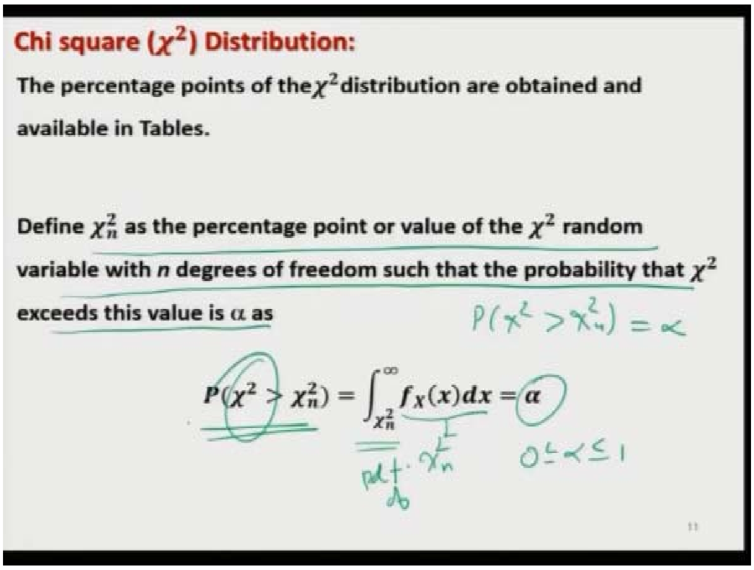
Now, if you want to find out different types of probabilities, earlier when people were not using the software some chi squared probabilities were compiled in a tables and they were called as chi square probability tables. So just for information you must know these things because if you are using it from the book then this chi square tables are directly available over there and well we can also compute them very easily in the R software also.

So, we define this chi square n as the percentage point or the value of the chi square random variable with n degrees of freedom such that probability that chi square exceeds this value, probability that like this one chi square exceeds chi square n and suppose this value here is alpha, some probability value so obviously this alpha is going to lie between 0 and 1. So you can see here this probability can be computed by the integral chi square n to infinity and then fx dx where fx is the probability density function of chi square n.

(Refer Slide Time: 19:34)

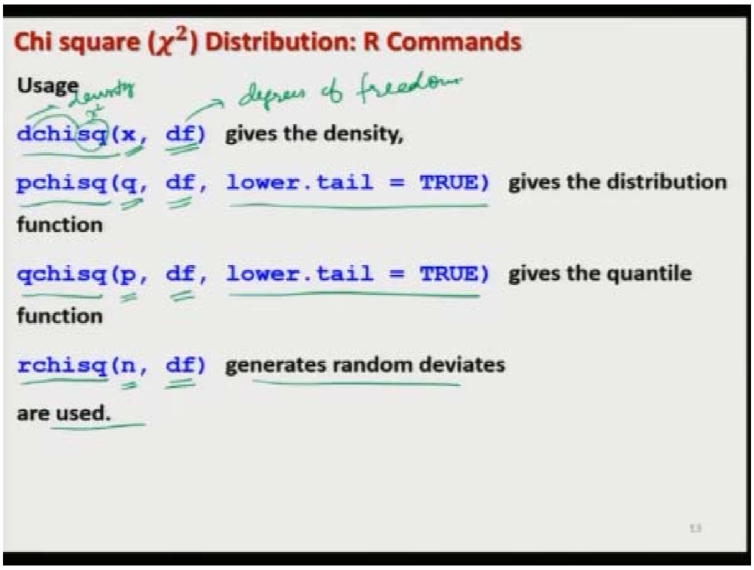


n	α = .995	α = .99	α = .975	α = .95	α = .05	α = .025	α = .01	α = .005
1	.0000393	.000157	.000982	.00393	3.841	5.024	6.635	7.879
2	.0100	.0201	.0506	.103	5.991	7.378	9.210	10.597
3	.0717	.115	.216	.352	7.815	9.348	11.345	12.838
4	.207	.297	.484	.711	9.488	11.143	13.277	14.860
5	.412	.554	.831	1.145	11.070	12.832	13.086	16.750
6	.676	.872	1.237	1.635	12.592	14.449	16.812	18.548
7	.989	1.239	1.690	2.167	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	23.685	26.119	29.141	31.319
15	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]



So, if you try to solve it you can get, but you see the solving such things is very difficult, so we have this type of table I am just giving you here a snapshot that here there are various values of alpha which are given here, which are given here like this, α equal to 0.995, 0.99 and so on and there you can see here are the degrees of freedom. So, for example, if you want to know the value of chi square with 5 degrees of freedom and alpha is equal to 0.995, so you have to look here this is the 5 degree of freedom, this is here alpha equal to 0.995 and then you can see here corresponding to this, this is the value 0.412. So, what is this value this is here this value.

(Refer Slide Time: 20:16)

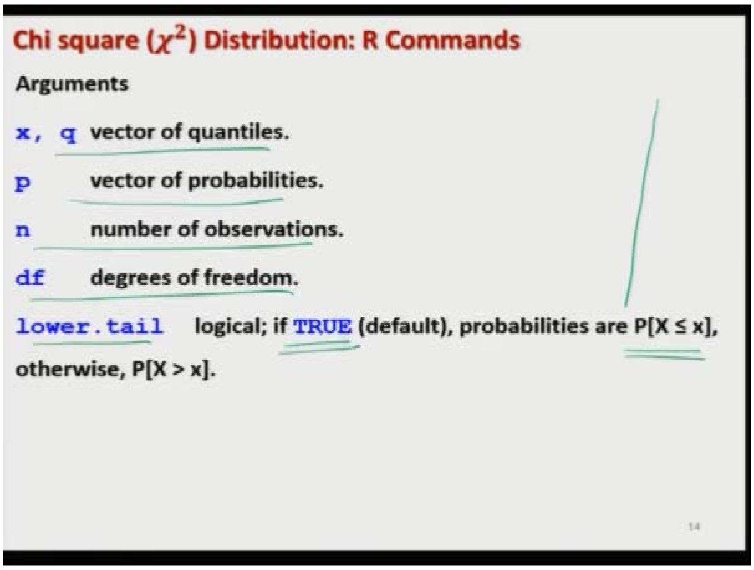


And now I would try to show you that how you can use it in the R software that is very simple now. So, if you want to find out the density or you want to compute the density you have to use the command dchisq, so this is dchisq, so d for density, chi for this chi square like this. So, you have to simply give here the value here x and the degrees of freedom so df is going to indicate the degrees of freedom, and this degree of freedom will be changing depending on the given conditions.

And similarly if you want to find out the CDF of this chi square distribution then we have the command here pchisq as earlier, then you have to give here the value at which you want to find out this CDF the degrees of freedom by the parameter df and then you have to give here lower dot tail is equal to TRUE or FALSE depending on your requirement, now you know how to use it.

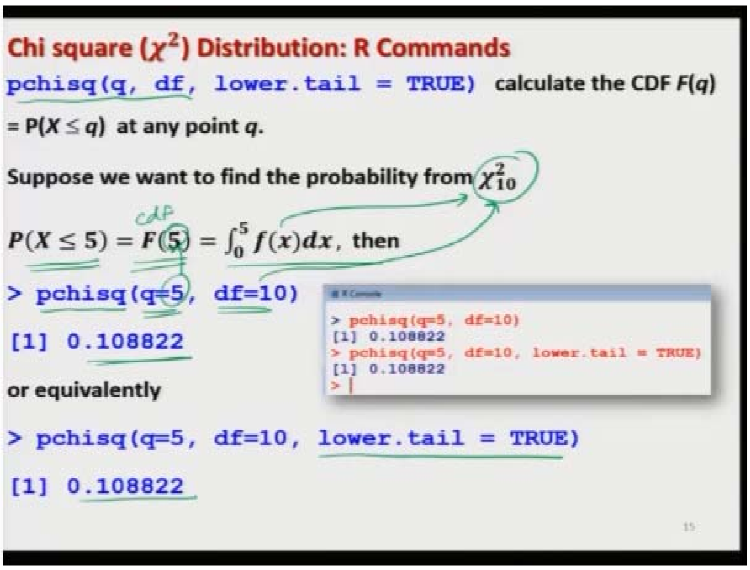
Similarly, if you want to find out the quantiles, then you have to give the command here qchisq and then you have to give the value for which you want to find out the quantile then degrees of freedom and then lower dot tail is equal to TRUE or FALSE depending on your requirement and similarly if you use here the command rchisq this will generate the random numbers, so if you try to give it here n this is the total number of random numbers that you want with the degrees of freedom. So, you can see here in this commands also the probability density function of chi square is going to be controlled by the parameter df that is degrees of freedom and that is what I was trying to explain you earlier.

(Refer Slide Time: 22:08)



And these are here the simple commands that you know that x and q are the vector of quantiles, p is the vector of probability, n is the number of observation. df is the decrease of freedom and lower dot tail will give you the probability X greater than equal to x, if it is TRUE and FALSE.

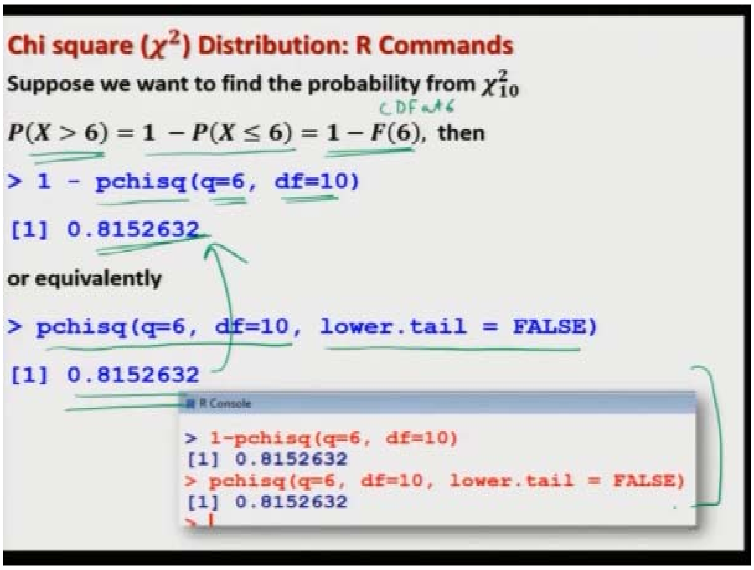
(Refer Slide Time: 22:26)



Now, let me try to take some example suppose, I want to compute the probability from the chi square distribution with 10 degrees of freedom suppose, I want to compute the probability that X is less than equal to 5 that we know that this is your CDF F at value 5. So that can be obtained by the $\int_0^5 f(x)dx$ where the f(x) is going to follow a chi square with 10 degrees of freedom, you can use the chi square probability tables also but that was an older method, now we are going to use the R software.

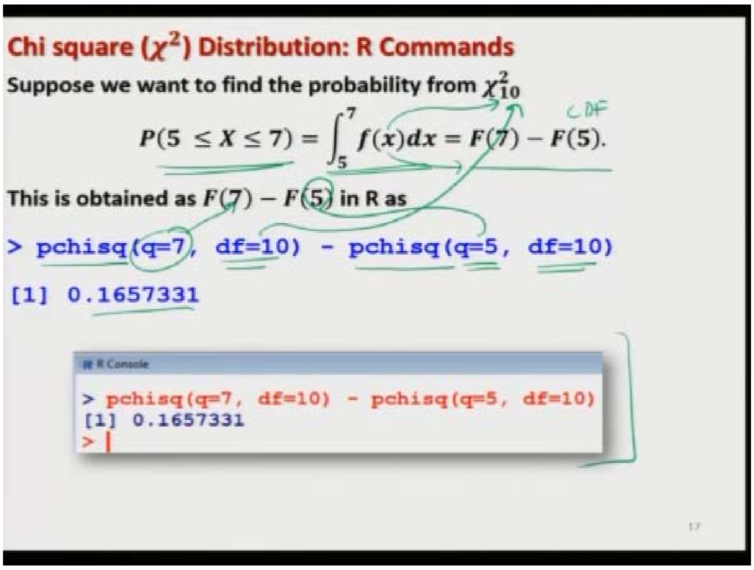
So, for that we have a command here pchisq that we know, so you have to simply give here pchisq, q is equal to here 5, so this 5 is coming from here and then you have to give here the degrees of freedom equal to 10 which is coming from here by the specification of the chi square distribution and R will give you this value 0.108822. And if you want to use here the command or the option in the command as lower dot tail is equal to TRUE, that is the default option it will give you the same value 0.108822.

(Refer Slide Time: 23:38)



Now, I try to take one more example and we want to compute the probability X greater than 6, so that can be expressed as 1 minus probability of X less than equal to 6 that is simply over here 1 minus F(6) which is the CDF, at 6. So this probability can be computed by writing 1 minus pchisq, q is equal to 6 and df equal to 10 which will come here like as 0.8152632 and if you do not want to use this idea of 1 minus CDF then you can directly use here this option that lower dot tail is equal to FALSE and you write the same thing pchisq, q is equal to 6 and df equal to 10 and it will give you this value 0.8152632 exactly in the same that you obtained earlier here.

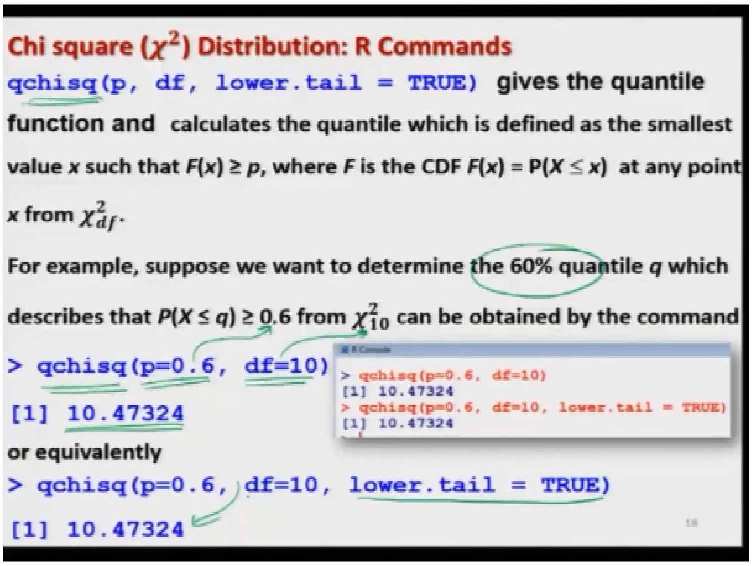
(Refer Slide Time: 24:35)



And this is here the screenshot I will try to show you it on the R console also. Similarly, if you try to compute the probability like X is lying between 5 and 7 so that can be obtained by the $\int_5^7 f(x)dx$, where f(x) is going to follow a chi square distribution and this probability can be written that we know as F(7) - F(5) where F is your CDF.

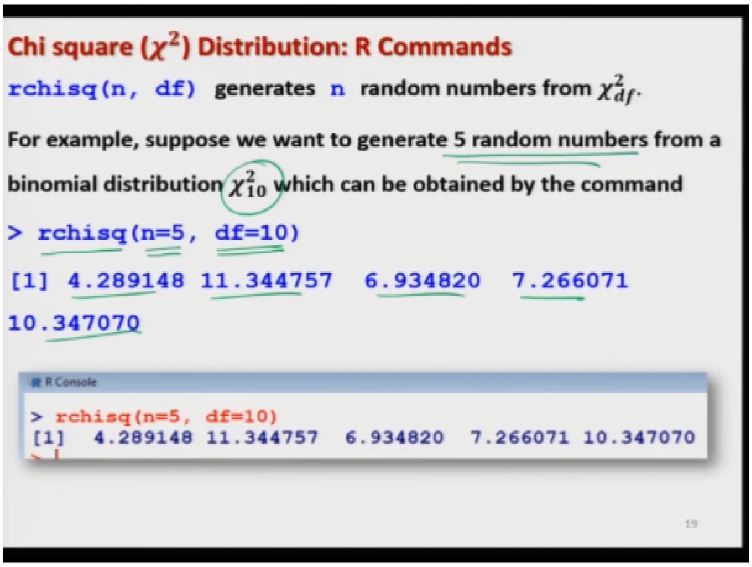
So this F(7) and F(5) can be obtained for this chi square distribution by the command pchisq, q equal to 7 coming from this 7 and then df equal to 10 coming from this distribution chi square and minus pchisq, q is equal to 5 and df equal to 10, so this 5 is coming from this 5, and if you try to solve it you will get here this value 0.1657331 and this is here the screenshot of the same operation, so you can see here the computing different types of probabilities is not difficult at all.

(Refer Slide Time: 25:30)



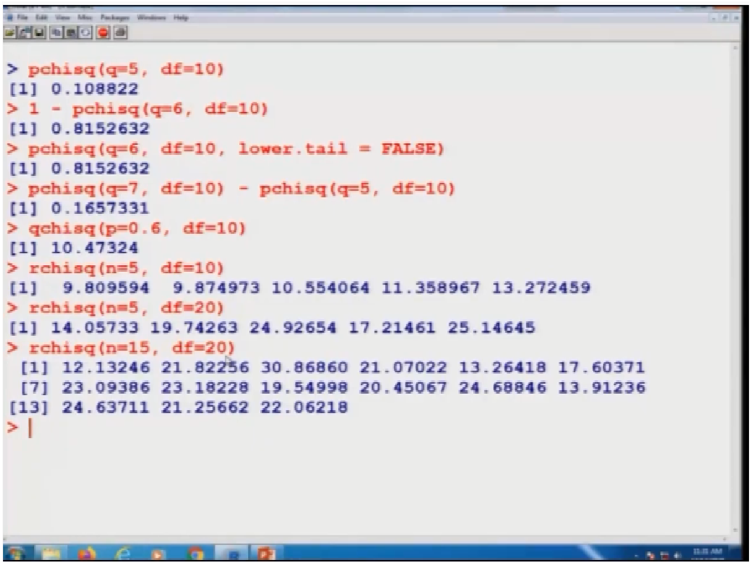
Similarly, if you want to compute the quantile from the same chi square distribution, so then we have a command here qchisq and suppose, I want to find out the 60 percent quantile, so for that we have to give here the qchisq, then I have to give here p is equal to 0.6 and df equal to 10, so the 0.6 is coming from here and df is coming from the specification of the distribution and you can see here this value comes out to be 10.47324. And well if you want to use here the option lower dot tail equal to TRUE, then once again it will give you the same value. So, it is not difficult at all.

(Refer Slide Time: 26:13)



And similarly, if you want to generate the random numbers from this chi square distribution suppose, I want to generate 5 random numbers from chi square distribution with 10 degrees of freedom, so that can be obtained by the command here rchisq, n is equal to 5 and df equal to 10 and you can see here 1, 2, 3, 4, 5 random numbers are generated and you can see that all the random numbers are greater than 0 and this is here the screenshot of the same outcome.

(Refer Slide Time: 27:00)



So, now look let me try to show you these operations on the R console also. So let me try to compute the probabilities which I have shown you here well it is very simple you simply have to put them in the R console and this chi square is built-in in the base software, so you can see here this probability is coming out to be 0.108822 this is the same probability here.

And similarly, if you want to compute this probability of probability greater than 6 you can see here this comes out to be here like this without any problem and if you try to use here this command and that lower dot tail is equal to FALSE then the same value will be obtained here you can see here this value and this value they are the same.

Similarly, if you want to compute here the probability that X is lying between 5 and 7 that we consider, you can see here this is the same probability that you have reported and similarly if you want to find out the quantiles you can see here you can find out the quantiles that is the 60 percent quantiles here like this 10.47324 and you can see here this is the same value that you had obtained earlier.

And similarly if you want to generate here the random numbers suppose, I want to generate the 5 random numbers then they are here like this and if you want to change here the degrees of freedom suppose, if I take a 20 you can see here this values are going to be quite different from the earlier one and if you try to generates instead of 5 you want to generate 15 random numbers, you simply have to give here n is equal to 15 and you can see here there are 15 values from chi square with 20 degrees of freedom.

So, you can see down that it is not really a very difficult thing to execute the chi square distribution, it is a computation of different types of probabilities, quantiles, random numbers etcetera, in the R software that is pretty simple and you know now you have done all the distributions and these commands are very similar to those things. So, now I would stop in this lecture but I will request you that you please try to look into the concepts of chi square statistics and you can see here that chi square is also the distribution of a statistics which is summation i goes from 1 to n N(0, 1) square.

So, whenever you will have random numbers $X_1, X_2, \ldots, X_n$ and if you try to square them, you try to sum them their distribution will follow simply chi square with n degrees of freedom, this is the main result of this lecture which you have to keep in mind. So, you try to revise this lecture and I will see you in the next lecture with one more sampling distribution that is t distribution till then good bye.