Statistics for Experimentalists

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

Lecture - 19
Chi-square distribution

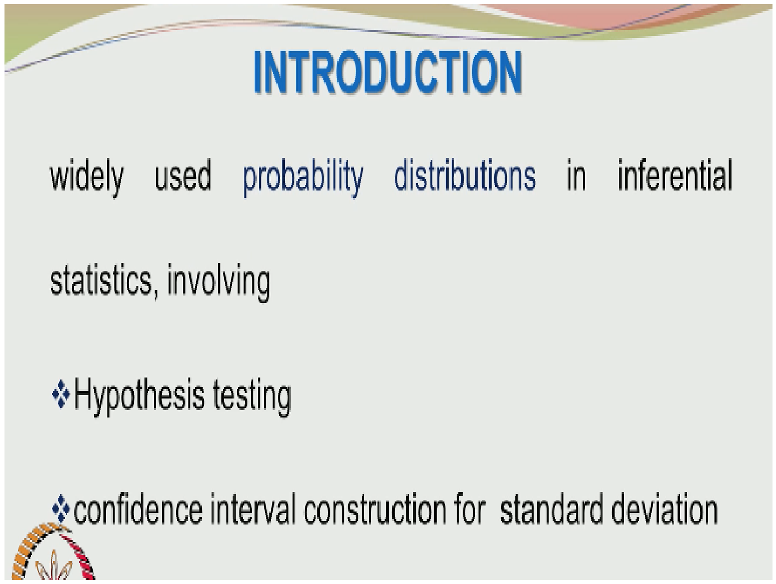
Welcome back we were looking at the T distribution now we will be looking at the Chi-square distribution. Why should we learn this interesting probability distribution function? We were so far talking about the population mean parameter mu equally important was the population parameter sigma namely the standard deviation. Since sigma is also not known we have to estimate it.

And we use the sample variance to provide the estimate for the population variance sigma squared. So, just as the population parameter mu we can talk about the estimation of the parameter sigma squared. We can also construct the 95% confidence interval for sigma squared based on the sample variance s squared or using the information contained in the random sample. We can construct 95% confidence interval.

How we can do it? well need the distribution of the sample variances be normal will it follow the T distribution these are the 2 distributions we have seen so far. The answer is a simple no the distribution of the sample variances does not follow the T distribution does not usually follow the normal distribution it follows the Chi-square distribution. So, we will be looking at the properties of the Chi-square distribution.

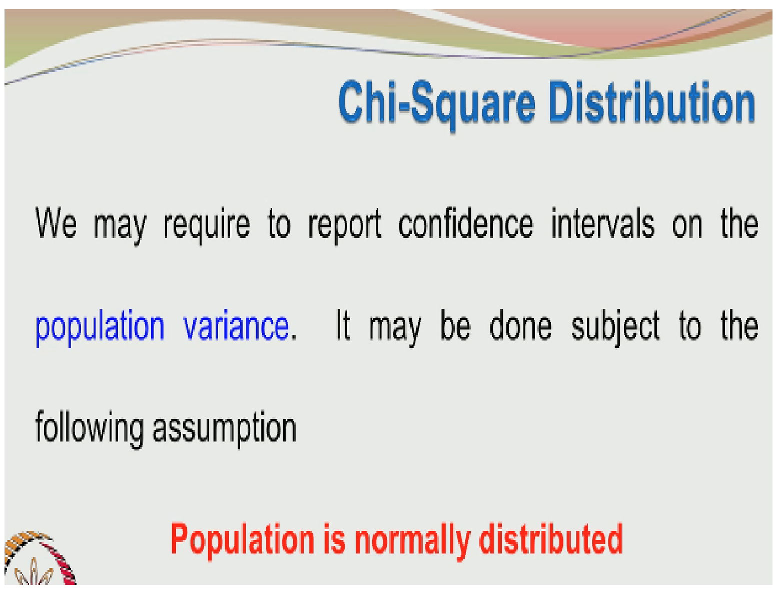
Some of the properties are very interesting we will see them shortly.

(Refer Slide Time: 02:19)



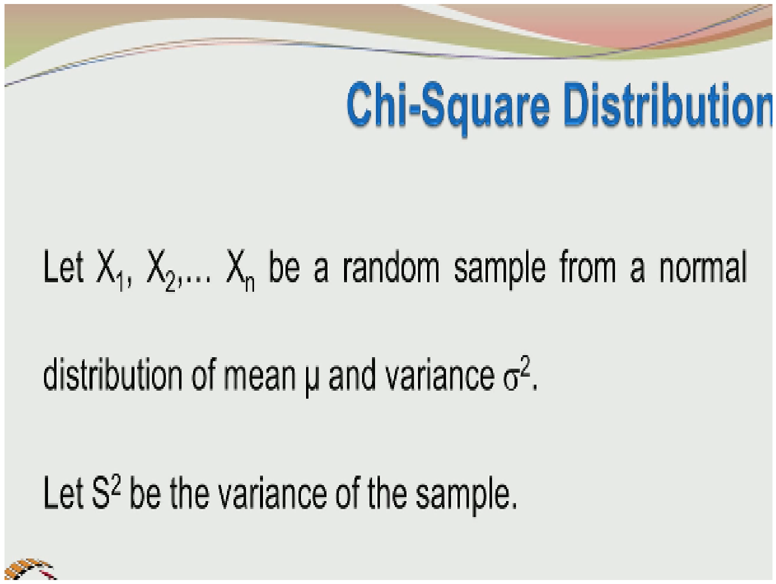
So, the Chi-square distribution finds application in hypothesis testing and confidence interval construction for the standard deviation.

(Refer Slide Time: 02:30)



We may require to report confidence intervals on the population variance sigma squared. The fundamental assumption the make is the population is normally distributed. This is not a very serious assumption as many populations do in fact tend towards the normal distribution and hence small deviations from this normality is not serious.

(Refer Slide Time: 03:01)



So, we are doing all this analysis based on a single sample of size n drawn from the population. So, we have x1 x2 so on to xn be the members of a random sample taken from a normal distribution of mean mu and variance sigma squared. We can calculate the sample variance S squared once the sample is drawn. So, until the sample is actually drawn the sample is an abstract entity we have a statistic S squared defined based on x1 x2 so on to xn.

These things we have covered in the previous classes. Now we have to define a particular random variable called the Chi-squared random variable. We have seen several random variables before then we also saw a standard normal variable for the random sample mean x bar. It was defined us x bar-mu/sigma/root n when the population standard deviation was not known, and the sample size was small then we resorted to the T random variable.

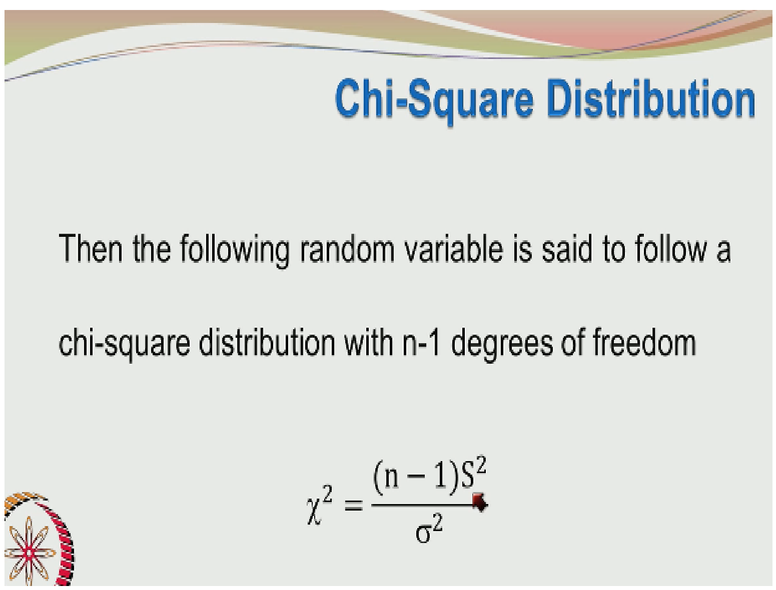
T was defined as x bar-mu/s bar/root n. Now we are talking about another population parameter namely sigma or sigma squared okay. We know that the sample variance s squared is an unbiased estimator of the population variance sigma squared. Now we are using the random sample variance to describe the population variance sigma squared. So, far we have been talking about the population parameter mu which was the represented by the sample mean x bar.

The sample mean x bar was also an unbiased estimator of the population mean mu s squared the sample variance is also an unbiased estimator of the population variance sigma squared. Now when you take many samples all these samples may have different sample variances the first sample may have a sample variance of 100 the second one may may have a value of 120 the third one may have 90 and so on.

So, which of these is a true representation of the actual unknown population variance sigma squared. Since all of them may be thought of as true representations because the sample variance is an unbiased estimator of the population variance sigma squared. However, we can ask which of these is lying closer to the actual population variance for which we have to follow the confidence interval concepts.

But before we go into all of them we now understand that there may be several sample variances, and these will follow a certain probability distribution. Okay the sample variance s squared is itself a random variable so it will follow a probability distribution. Now what probability distribution does the sample variance follow? it follows the Chi-square distribution for that we have to define the Chi-square variable.

(Refer Slide Time: 07:11)



As n-1 s squared/sigma square okay s squared is the sample variance and sigma squared is the population variance sigma squared we do not know s squared we know once we have drawn the sample. And n-1 is the sample size-1 it represents the degrees of freedom and it was used in the computation of s squared recall that the sample variance s squared=sigma xi-x bar whole square/ /n-1.

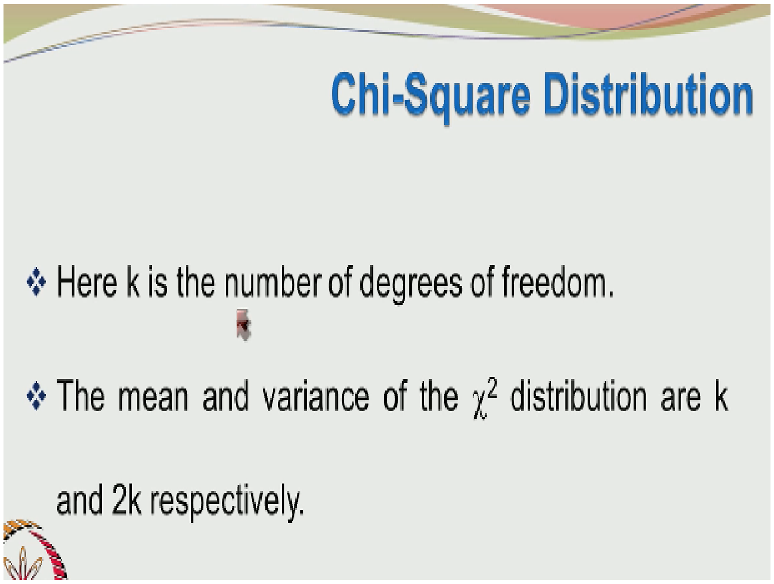
So, we may think of n-1 s squared as the sum of the square of the deviations from the mean. Deviations of what? deviations of the sample element attributes from the mean attribute value and that deviation is squared So, we have the Chi-square variable as n-1 s squared/sigma squared and this also depends upon the degrees of freedom. Let us see the probability distribution function for the Chi-square random variable.

We will not be using this probability distribution function extensively in our analysis. We will be using the probability table for the Chi-square distribution rather than actually computing the probabilities using this function. We know by now that minus infinity to infinity f of x dx=1 generally but in the present case the x value are always positive. So, this function is defined for x>0.

And so, we have 0 to infinity f of x dx should be equal to 1 that may be shown it is not required right now. It is beyond the scope of our current discussion. Here we have 1/2 power k/2 gamma k/2 x power k/2-1 e power-x/2. Here k is a parameter in the Chi-square distribution it matches with the degrees of freedom. We were talking about with respect to the sample variance calculation what is this gamma function? we have seen this gamma function pretty frequently.

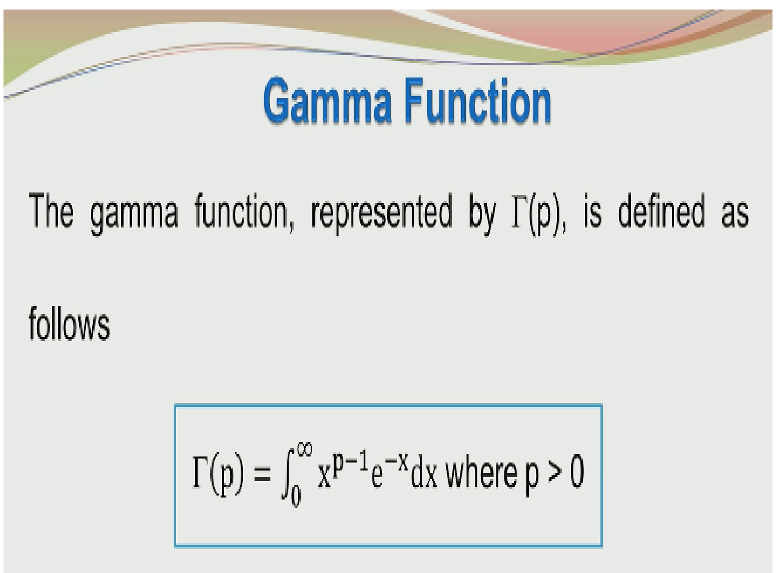
We saw the gamma function also in the description for the T distribution. So, I thought maybe I will just give a brief introduction to the gamma function as well.

(Refer Slide Time: 10:12)



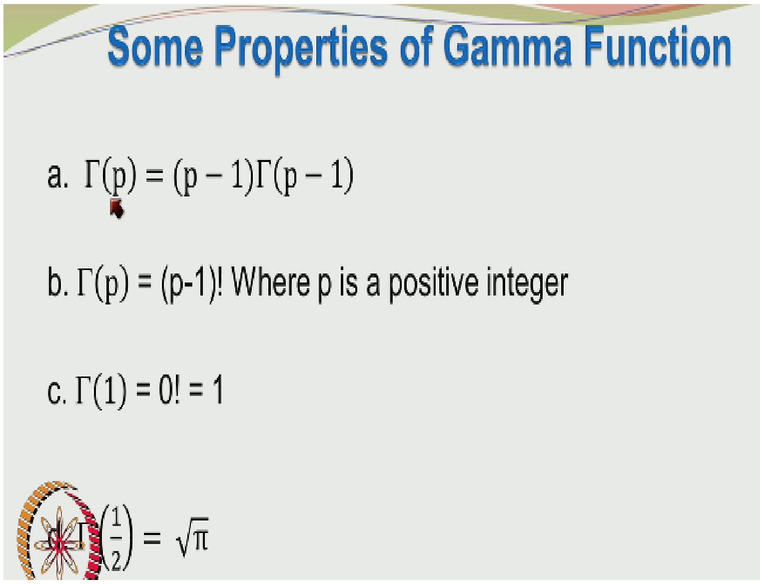
Before we do that, we will see what the mean and variance of the Chi-square distribution are the mean of the Chi-square distribution is k where k is the parameter referred to as the degrees of freedom. The variance of the Chi-square distribution is 2k.

(Refer Slide Time: 10:39)



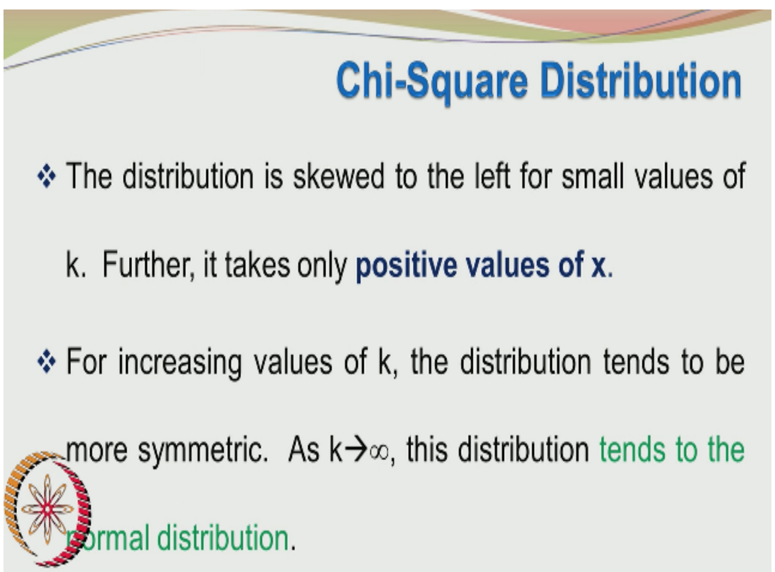
The gamma function is given by 0 to infinity x power p-1 e power-x dx where p is >0. So, this is the integral where p as the parameter.

(Refer Slide Time: 10:53)



So, it can be shown that gamma p=p-1 gamma p-1 kind of a recursive relationship and if this is carried out fully for a positive integer p we get gamma p=p-1 factorial gamma 1 is 0 factorial which is 1 and gamma 1/2 is root pi these are some important properties of the gamma function.

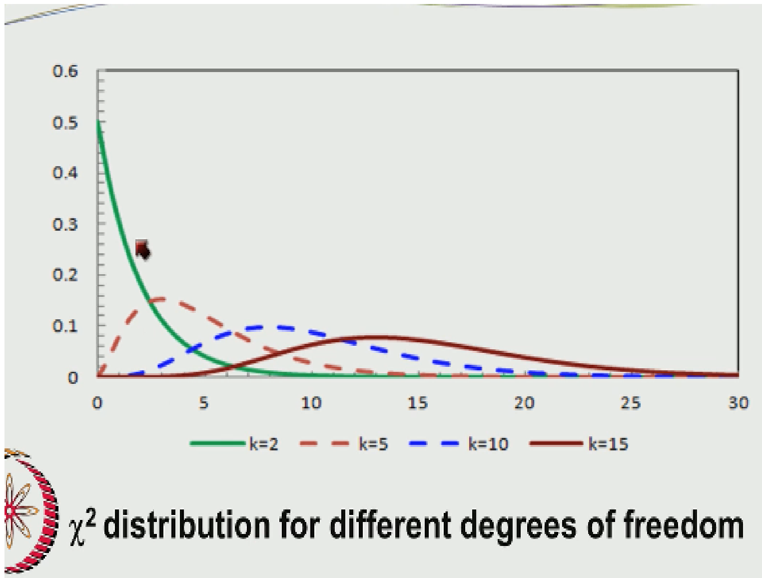
(Refer Slide Time: 11:22)



Coming back to the Chi-square distribution the distribution is skewed to the left in other words it is not symmetric the skewness is pronounced for small values of k and the distribution is defined only for positive values of x and we can see that as the degrees of freedom increase the distribution tends to more symmetric shape as k tends to infinity as a degrees of freedom becomes quite large.

The distribution tends towards the normal distribution again taken to this extreme this skewed distribution tends to normal distribution.

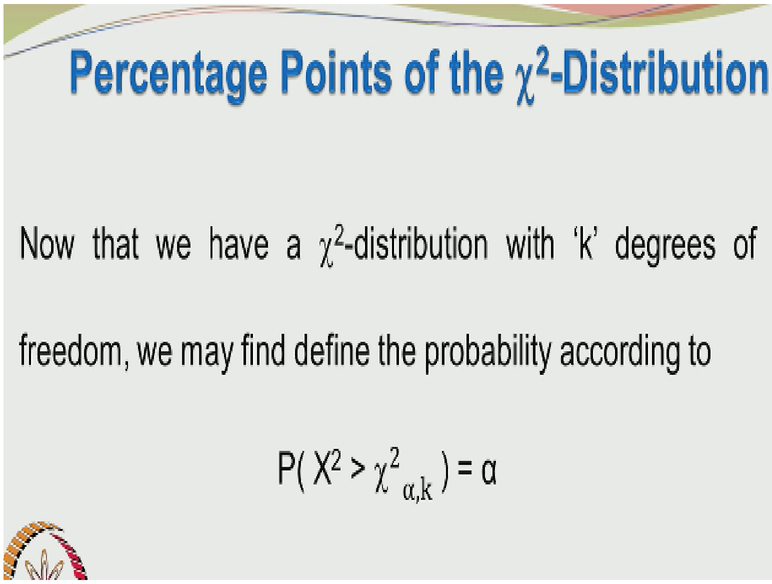
(Refer Slide Time: 12:23)



These are some typical representations of the Chi-square distribution generated with spreadsheet and this is the x value and f of x value is on the y axis. It can be seen that the distribution is quite skewed for a small k here it is 5 if k becomes very small k=2 then it becomes more of an exponential kind of distribution. And when you go for k=10 it sorts of becomes less skewed and for k=15 the spread the more but it is symmetric.

And as k tends to infinity it goes towards the normal distribution.

(Refer Slide Time: 13:20)

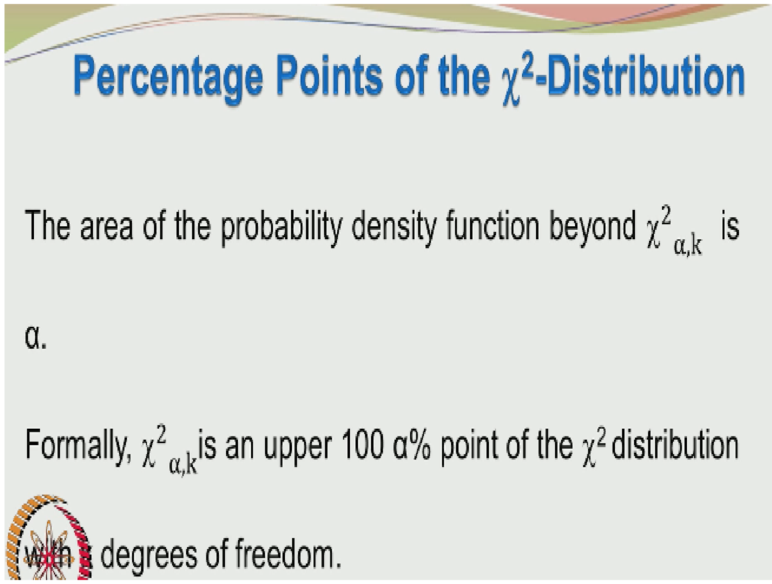


By now we should be familiar with these terminologies suppose we have a Chi-square distribution with k degrees of freedom we may define the probability according to probability of Chi-Square> a specified Chi-square alpha k. K is the degrees of freedom and alpha is chosen such that it is= the probability value. We have seen this alpha before in our discussion on confidence intervals and the T distribution.

Alpha may take values typically like.01.025.05.1 etc. So, we are again talking about the upper tail probability. What is the probability of the Chi-square distribution beyond for values beyond the Chi-squared alpha k point? Suppose you have a distribution curve the Chi-squared alpha k are all values along the x axis. What is the probability that the squared random variable will take a value > Chi-squared alpha k.?

And that value will be nothing but alpha so this alpha and this alpha match the Chi-squared I repeat represents a point on the x axis of the Chi-square distribution curve.

(Refer Slide Time: 15:00)



So, the area of the probability density function below the curve beyond Chi-squared alpha k is alpha. So, formally we may talk in statistical language that Chi-squared alpha k is an upper 100 alpha % point of the Chi-squared distribution with k degrees of freedom. So, you have alpha k, k refers to the degrees of freedom and alpha refers to the upper tail probability. So, Chi-squared alpha k is an upper 100 times alpha %.