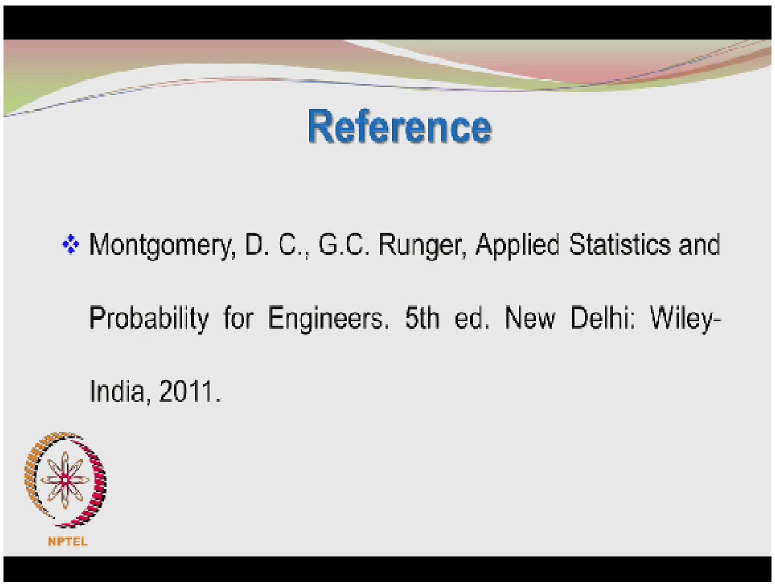
Statistics for Experimentalists

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

Lecture – 20
F-Distribution

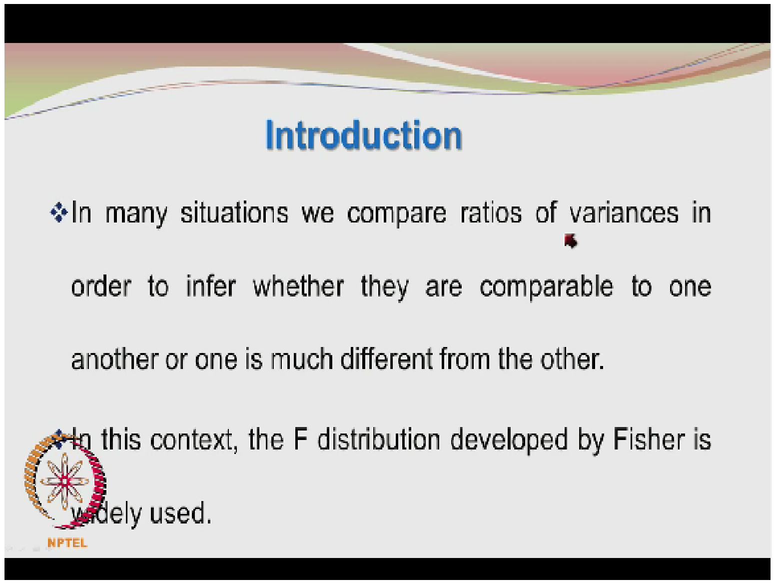
Hello everyone. In today's lecture, we will be looking at the F-distribution.

(Refer Slide Time: 00:21)



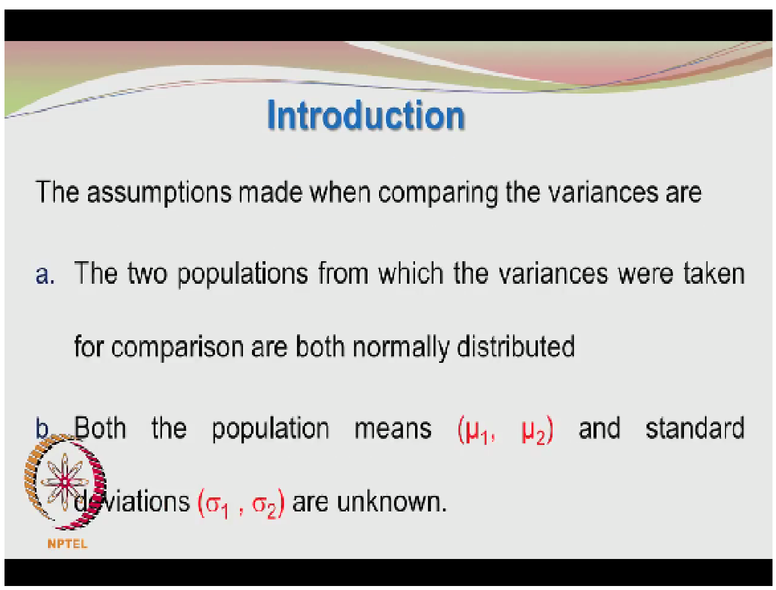
The reference book for those topic is the one written by Montgomery and Runger. Let me give you a brief introduction. So far we have been looking at the sample mean and the distribution of sample means. The variance of the population is also an important parameter and we want to also compare 2 variances and make decisions on them. So in this connection, the F-distribution is applied.

(Refer Slide Time: 01:13)



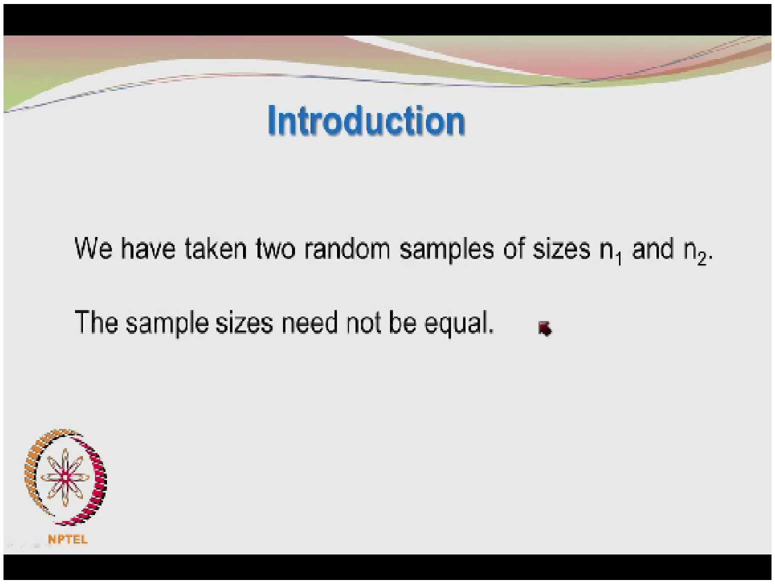
Here we compare ratios of variances in order to infer whether they are comparable to one another or one is much different from the other. The F-distribution developed by Fisher is widely used for this purpose.

(Refer Slide Time: 01:33)



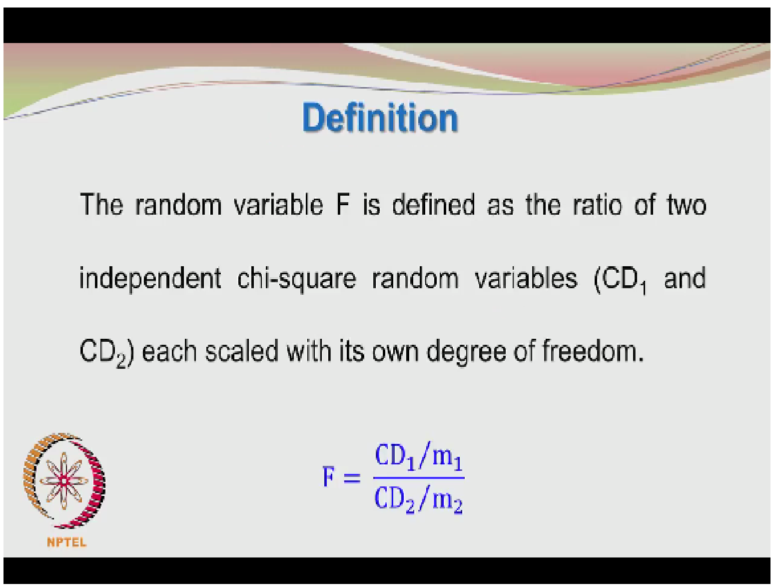
What are the assumptions made? The 2 populations from which the variances were measured for comparison are both normally distributed. The population parameters mu1, mu2 and standard deviations sigma1, sigma2 are not known. Let us assume that we have taken 2 random samples of sizes n1 and n2. The sizes need not be equal.

(Refer Slide Time: 02:16)



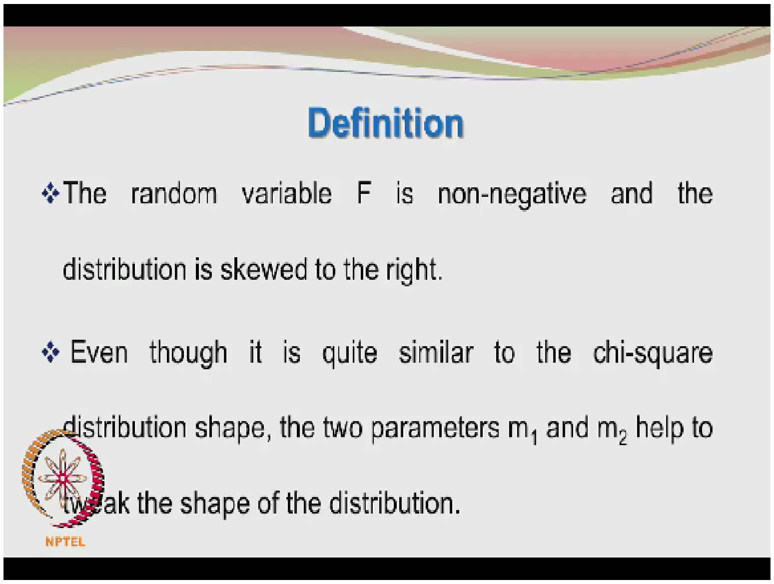
In other words, n1 need not be = n2.

(Refer Slide Time: 02:19)



Random variable F is defined as the ratio of 2 independent chi-square random variables CD1 and CD2, each of it being scaled by its associated degrees of freedom. So F=the chi-square distribution 1/m1, where m1 is the degrees of freedom associated with the chi-square distribution 1. If the second chi-square distribution is represented by CD2, then we scale it by its associated degrees of freedom namely m2. So we define the F random variable as CD1/m1/CD2/m2.

(Refer Slide Time: 03:16)

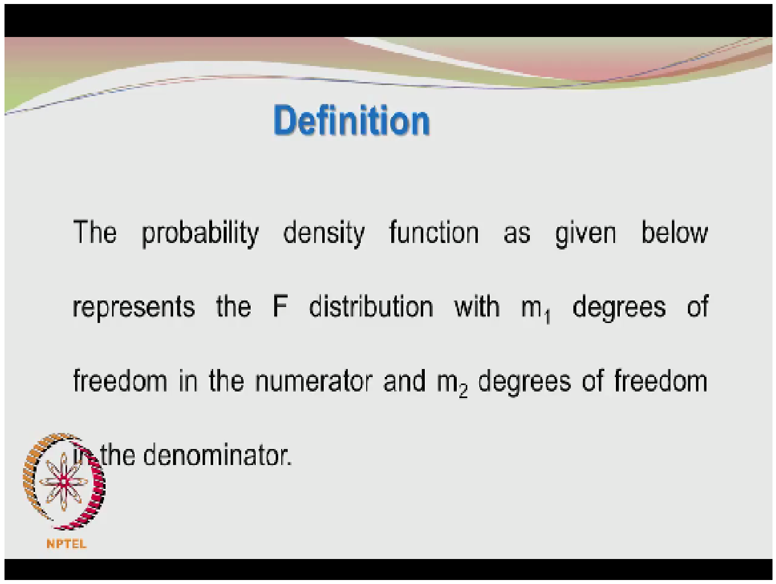


The random variable F is nonnegative and the distribution is skewed to the right. So the probability distribution function when plotted on the graph does not give usually a symmetric curve. It is giving a curve that is skewed to the right. Even though it is quite similar to the chi-square distribution shape, the 2 parameters m1 and m2 help to tweak the shape of the distribution.

We have seen both in the case of the T-distribution as well as in the case of the chi-square distribution, the degrees of freedom K was the parameter. In the F distribution also, we have 2 parameters, m1 and m2. If you recollect, the chi-square distribution 1 with its associated degrees of freedom, m1 was divided by chi-square distribution 2 with this associated degrees of freedom m2. So we can say the F-distribution has 2 parameters, m1 and m2 where ma and m2 are the degrees of freedom of the first chi-square distribution and the second chi-square distribution respectively.

So these 2 parameters m1, m2 may be changed to change the shape of the distribution. In certain cases, you may want to fit a probability distribution to your experimental data to see from which family of populations your experimental data is more fitting. So when you have an experimental trend, you want to fit up distribution to it and if you have 2 parameters, then you have a more flexibility or more possibilities of fitting nicely the curve to the experimental data points. So m1 and m2 help to tweak the shape of the distribution.

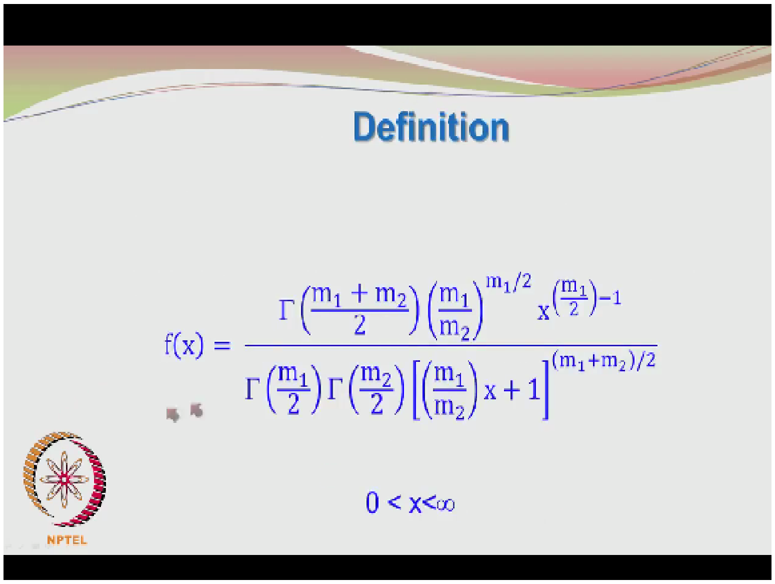
(Refer Slide Time: 05:47)



So again as usual, we will show the mathematical form for the probability density function and that would represent the F-distribution with m1 degrees of freedom in the numerator and m2 degrees of freedom in the denominator. So we are having m1 as degrees of freedom in the numerator, m2 as the degrees of freedom in the denominator. So do not try to take m2 to the top and m1 to the bottom and say m2 is the numerator degrees of freedom, it is not like that.

You are focusing on the chi-square distribution 1 which is having m1 degrees of freedom and chi-square distribution 1 is present in the numerator and similarly you have chi-square distribution 2 which is present in the denominator and we talk of m2 degrees of freedom in the denominator.

(Refer Slide Time: 06:50)

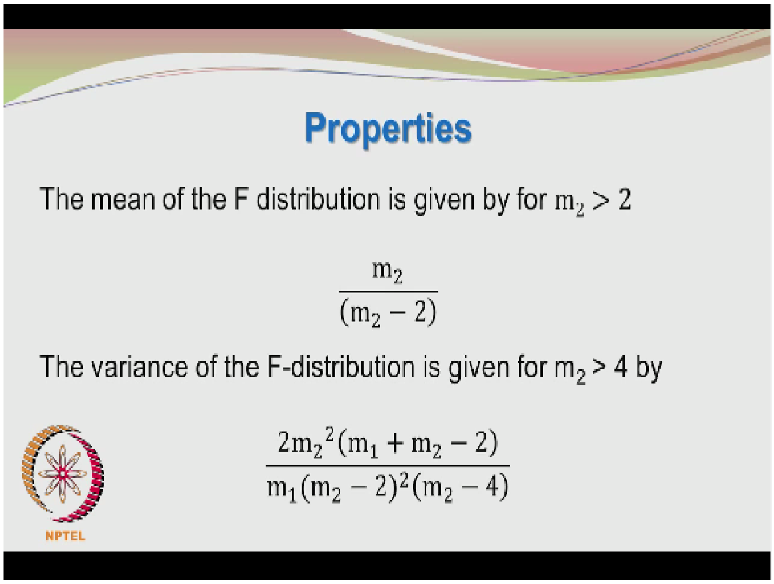


So this is an impressive or difficult looking probability distribution function. It depends on how you want to look at it. Fortunately, or unfortunately, we will not be really using this distribution in our calculations of probabilities. We would be rather using the probability charts for this purpose; however, it is useful to see the shape of the distribution and also the mathematical form of the distribution.

So this is again the gamma function in connection with the chi-square distribution, I gave a brief introduction to the gamma function. So here you have gamma of m1+m2/2. Obviously m1 and m2 must be integers and they may be 3 and 2 for example or 5 and 7 or it may be 8 and 5. So what I am trying to say is here you may have gamma 6.5 or gamma 4.5 and values for such nonnegative numbers do exist.

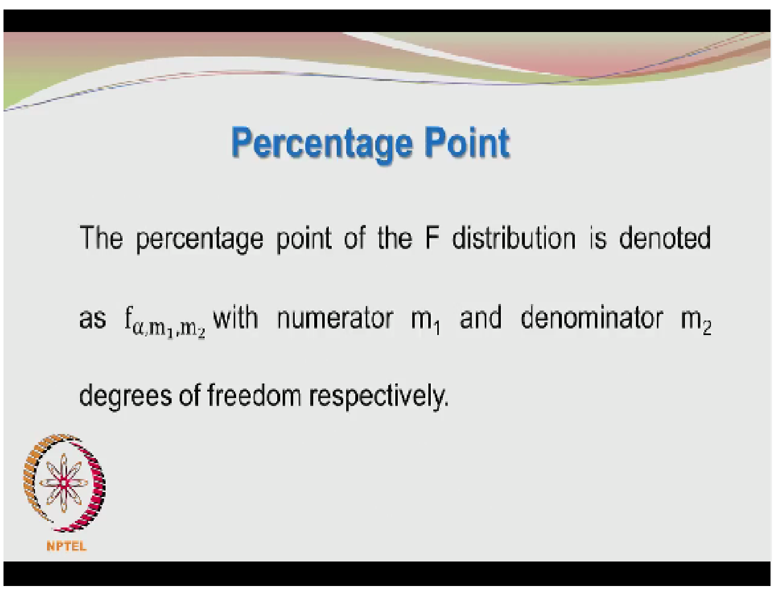
So m1 and m2 must be integers and obviously you cannot have a degree of freedom of 0. The degree of freedom should be at least 1 and another important thing to note here is the independent variable x, it is present twice. One in the numerator here and another in the denominator and it can take only positive values. So the Fisher distribution is describing the ratios of 2 variances. The variances themselves are positive quantities. So here we have x taking only positive values ranging from 0 to infinity.

(Refer Slide Time: 09:05)



The main for the F-distribution is given by m2/m2-2 and the variance of the F-distribution is given by 2m2 squared*m1+m2-2/m1*m2-2 whole squared/m2-4. So to ensure that these parameters do not blow up, we have to make sure that the degrees of freedom m2 is > 2 as far as the mean is concerned and for the variance, m2 should be > 4.

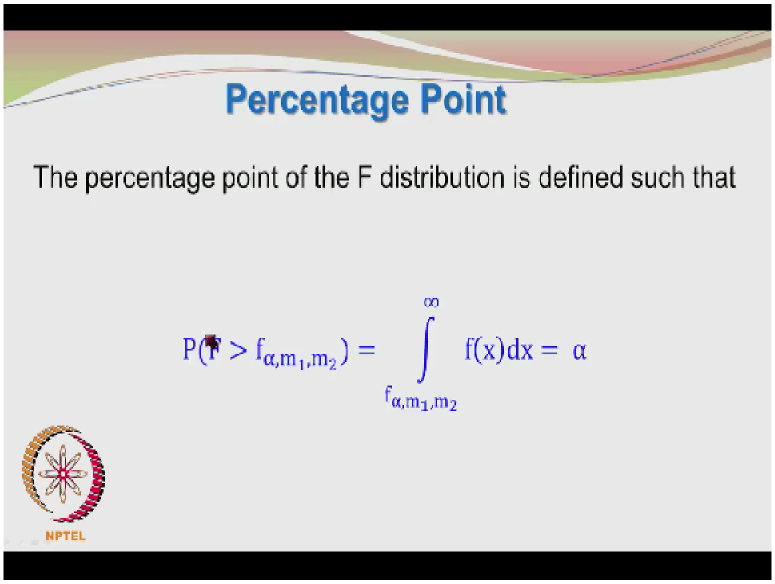
(Refer Slide Time: 09:41)



Now we are going to look at the percentage point of the F-distribution. We have already seen the percentage point for the standard normal variable. Also for the T random variable and the chi-square random variable. On similar lines, we are going to define the percentage point for the F random variable.

The percentage point of the F distribution is denoted as f of alpha m1m2 with numerator m1 and denominator m2 degrees of freedom respectively. So alpha is the level of significance. We were using it to define the confidence interval and alpha was typically taking values of point 0.01, 0.025, 0.05, etc. So the sequence is alpha must come first, then the numerator degrees of freedom and then the denominator degrees of freedom.

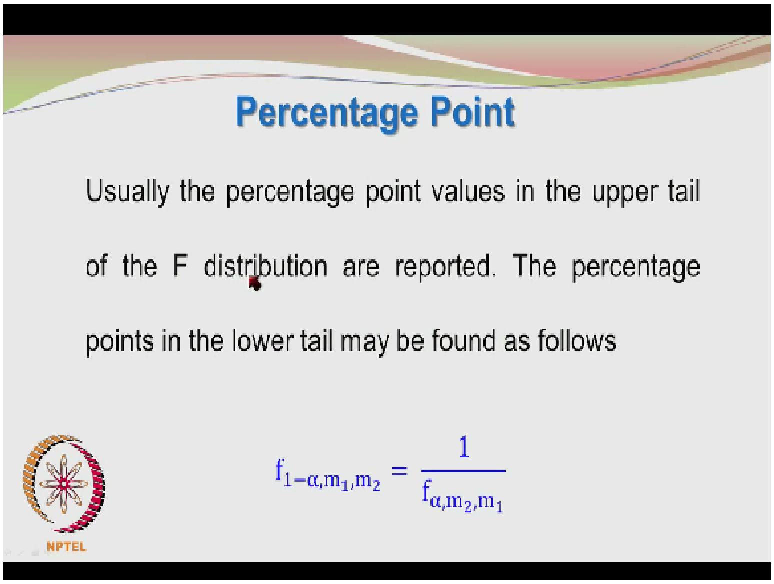
(Refer Slide Time: 11:01)



The percentage point of the F distribution is defined such that probability of the F random variable > f of alpha m1m2=the same value given as the lower limit, it is a numerical value, so that value comes as the lower limit. The upper limit is infinity and then f of xdx, that is = alpha. So define a number f of alpha m1m2 such that when it is used in the lower limit.

And then the probability density function which we saw a couple of slides back is incorporated here and the necessary integration is carried out, we get alpha. So it is a kind of an inverse problem. What is the F alpha m1m2 such that the probability of the Fisher random variable exceeding this number is alpha.

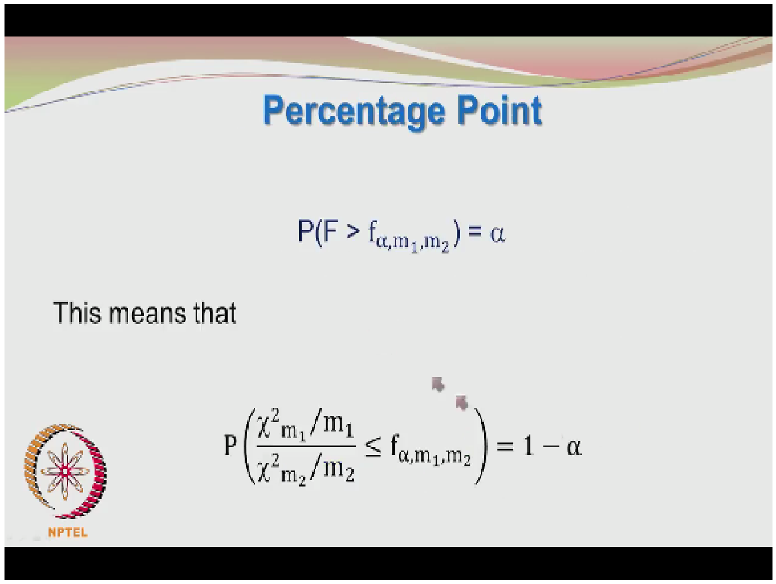
(Refer Slide Time: 12:28)



So we always talk on the percentage point values in the upper tail of the F distribution. The percentage points in the lower tail is given by f of 1-alpha m1m2. You have a distribution. You locate a point anywhere between 0 to infinity. Then what happens is, the area under the curve beyond the f alpha m1m2 will be alpha and then, okay. So that constitutes the upper tail region. Since the total probability = 1, the area under the curve below f of alpha m1, m2 will be 1-alpha.

So the probability is 1-alpha in the lower tail region if it is alpha in the upper tail region. Usually f of alpha values are reported. If we want to have the lower tail value f of 1-alpha m1m2, it can be shown that f of 1-alpha m1m2=1/f of alpha m2, m1. So what we have to note here is, we are changing 1-alpha to alpha, then we go from numerator to denominator and the sequence of the degrees of freedom also gets interchanged. It is m1m2 originally generally, then it becomes m2m1.

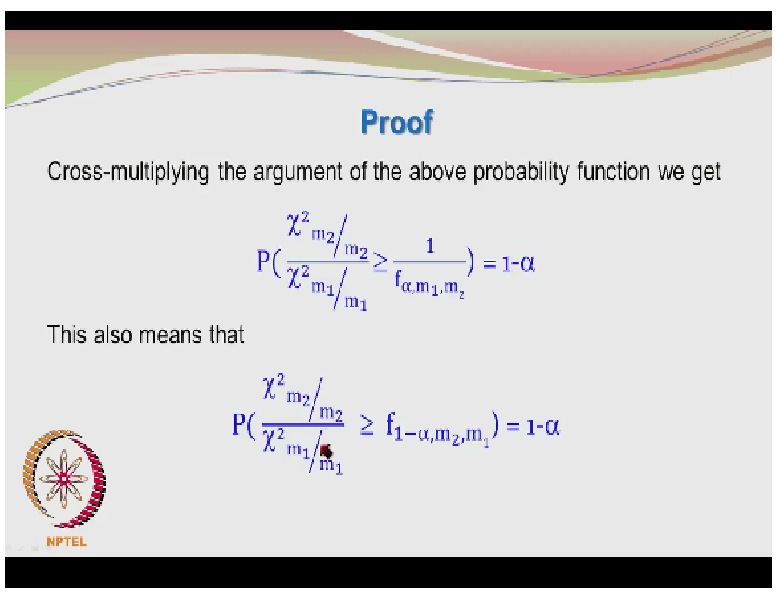
(Refer Slide Time: 14:21)



So we defined the f of alpha m1m2 such that or we identify the number f of alpha m1m2 such that probability of f>f of alpha m1m2=alpha. What does this mean? It means probability of chi-square m1/m1/chi-square m2/m2 is <= f of alpha m1m2=1-alpha. So the first expression defines the upper tail region. The second where I am substituting the definition for f here in terms of the 2 chi-square distributions, I am having it as probability of chi-square m1/m1/chi-square m2/m2 is <= f of alpha m1m2 and that is = 1-alpha, the lower tail probability or area under the curve.

We can cross multiply. So we can take chi-square m2/m2 here, chi-square m1/m1 below. Then the inequality sign will change and then you have 1/f of alpha m1m2. Let us see what happens?

(Refer Slide Time: 15:58)

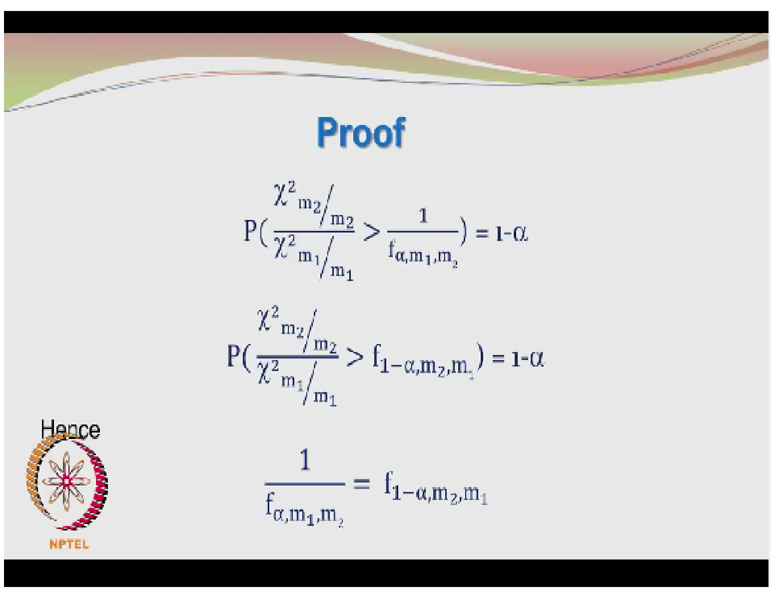


This is what we will get. If you want to check it up, please do it separately after pausing. So here we have probability of chi-square m2/m2/chi-square m1/m1 (()) (16:18) 1/alpha m1m2=1-alpha. Here the subscript is m2 and the subscript is m1 corresponding to the degrees of freedom associated with the chi-square distributions. Here m2 and m1 are the degrees of freedom, okay and you have to be a bit careful here because m1 and m2 are numerator and denominator degrees of freedom.

And you define the upper tail probability and even though now you get chi-square m2/m2, chi-square m1/m1, please do not lose track of the degrees of freedom and that was shown to be >= 1/f of alpha m1, m2 and that is = 1-alpha. So this f random variable. This also means that this random variable is >= f of 1-alpha m2, m1=1-alpha. So here also we are defining f of 1-alpha m2m1 such that the upper tail probability = 1-alpha.

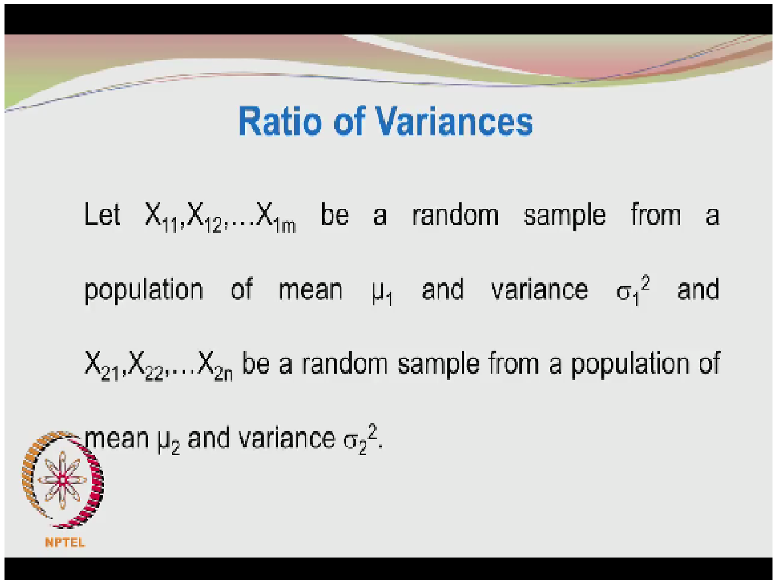
So both of them are the same. Here also it is 1-alpha, here also it is 1-alpha. We are talking about the same f random variable and hence this value should be equal to the value shown below. In other words, f of 1-alpha m2m1=1/f of alpha m1m2. Quite a nice derivation.

(Refer Slide Time: 18:47)



So f of 1-alpha m2m1=1/f of alpha m1m2.

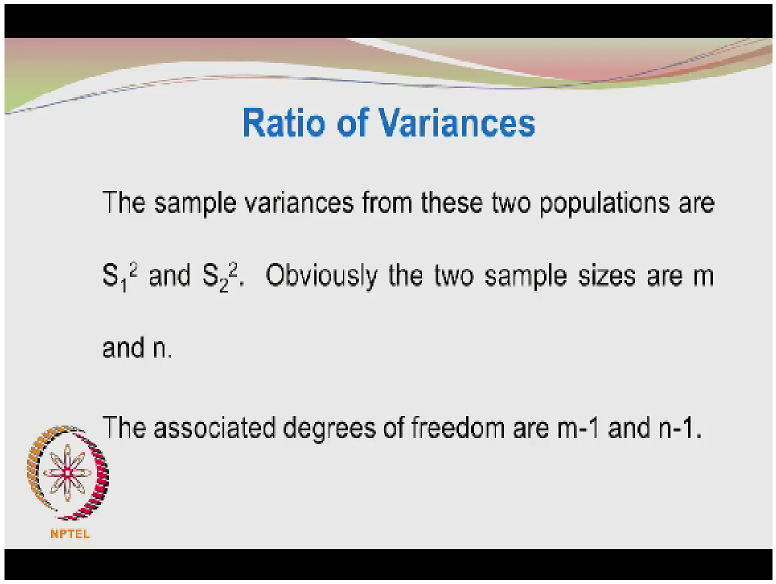
(Refer Slide Time: 28:54)



Now let us look at the ratio of variances. Let us take 2 random samples and we will denote the random variable X as Xij, i standing for the random sample index and j standing for the elements of the random sample i. So i will take values of 1 to 2 and j will take values of m and n. Let X11, X12 so on to X1m be the members of a random sample from population of mean mu1 and variance sigma1 squared.

If you want, you can say that I have taken this random sample comprising of m entities from population 1. So if you chose population 2 and take a random sample from the population 2 of size n, we call the random variables sampled as X21 second population first random variable, X22 so on to X2n and of course here, n is the sample size taken from the second population and the second population has mean mu2 and variance sigma2 squared.

(Refer Slide Time: 20:43)

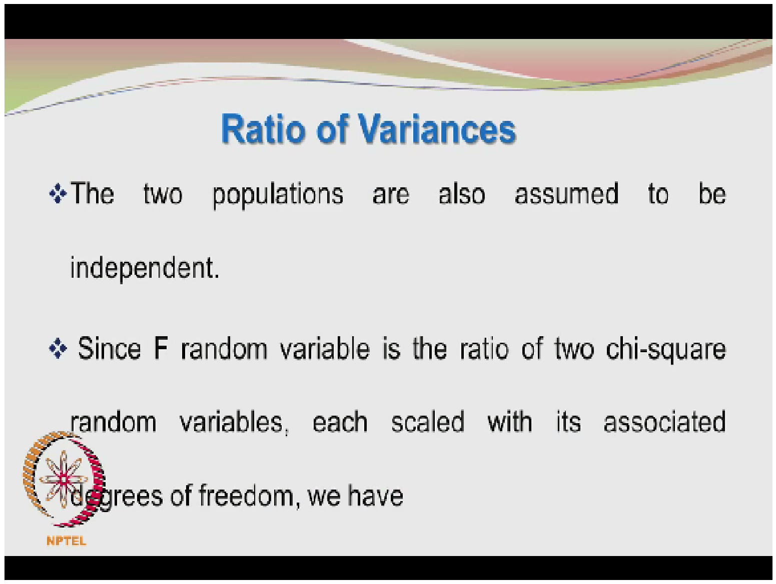


The sample variances may be computed in addition to the sample means. Now we are focusing more on the sample variances. We do not really use the mean of the random samples in our analysis here. This is in contrast to the analysis we were doing in the case of the distribution of the means, for example when we set up the confidence interval, we used, in certain situations, the sample variance S squared but now in our F distribution analysis, we are not talking that much about X bar, the random sample mean.

But we are talking mostly about S1 squared and S2 squared. In fact, so far in our discussion, X1 bar or X2 bar did not figure in them. So we have S1 squared and S2 squared as the sample variances from the 2 random samples we have chosen. So we have size m and size n. So the associated degrees of freedom are m-1 and n-1. We chop of 1 from the degrees of freedom because not all these squared deviations from the mean are independent.

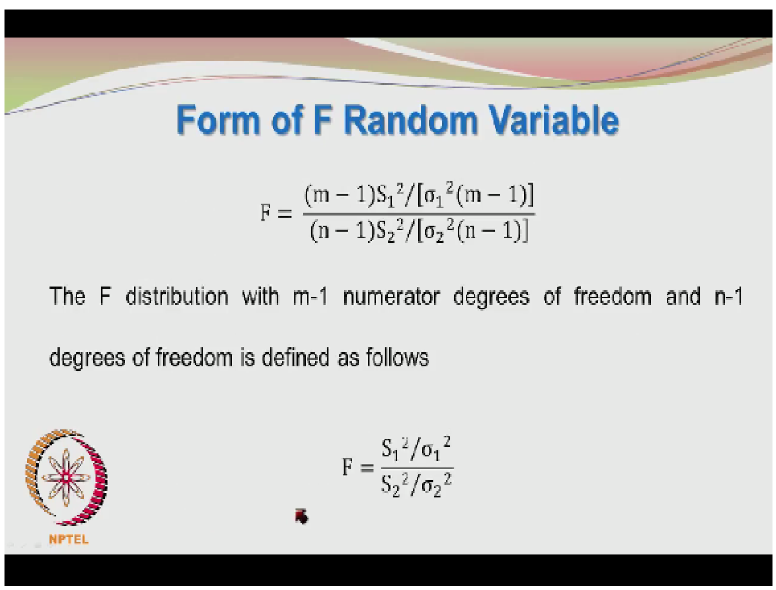
Only m-1 and n-1 independent squared deviations exist. Of course, you use the sample mean implicitly because when you want to calculate the sample variances, S1 squared and S2 squared, we use X1 bar and X2 bar. So implicitly, the sample mean is necessary but explicitly, it is not appearing in these calculations. The 2 populations are also assumed to be independent.

(Refer Slide Time: 23:04)



Since the F random variable is the ratio of 2 chi-square random variables, each scaled with this associated degrees of freedom, we have F=m-1S1 squared/sigma1 squared that refers to the first chi-square distribution that we are dividing by m-1.

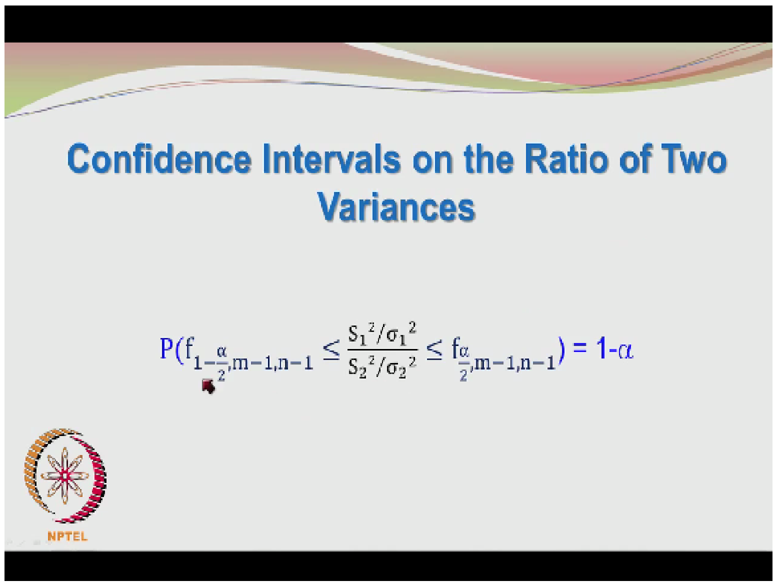
(Refer Slide Time: 23:15)



Next we are going to the denominator. We are talking about the chi-square distribution in the denominator that is given by n-1S2 squared/sigma2 squared and that we are scaling by the degrees of freedom associated with the second chi-square distribution and that is n-1. Obviously this m-1 and m-1 will cancel, n-1 and n-1 will also cancel, so we are simply left with S1 squared by sigma1 squared/S2 squared/sigma2 squared.

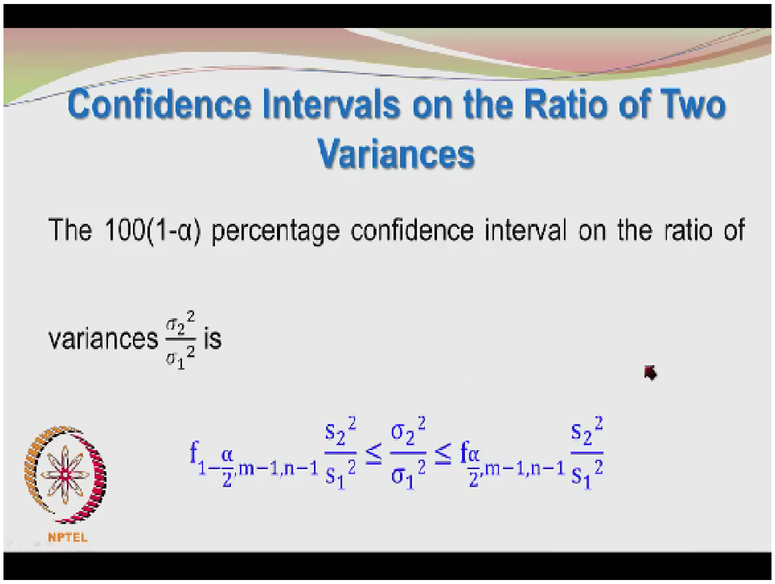
We may write it in a simple and compact fashion as S1 squared/sigma1 squared/S2 squared/sigma2 squared. Now we can develop confidence intervals on the ratio of 2 variances. What are the upper and lower limits that bound the ratio of 2 variances? Remember both sigma1 squared and sigma2 squared are not known. They are the first and second population variances respectively. We do not know them and so we have to consider the confidence interval for the ratio of the 2 unknown variances.

(Refer Slide Time: 24:59)



In the usual manner, we define probability of f of 1-alpha/2 m-1 n-1 is <= S1 squared/sigma1 squared/S2 squared/sigma2 squared is <= f of alpha/2 m-1 n-1, that probability = 1-alpha. So we have to identify 2 limits, the lower limit here and the upper limit here such that the probability of the f random variable lying between these 2 limits = 1-alpha.

(Refer Slide Time: 25:43)



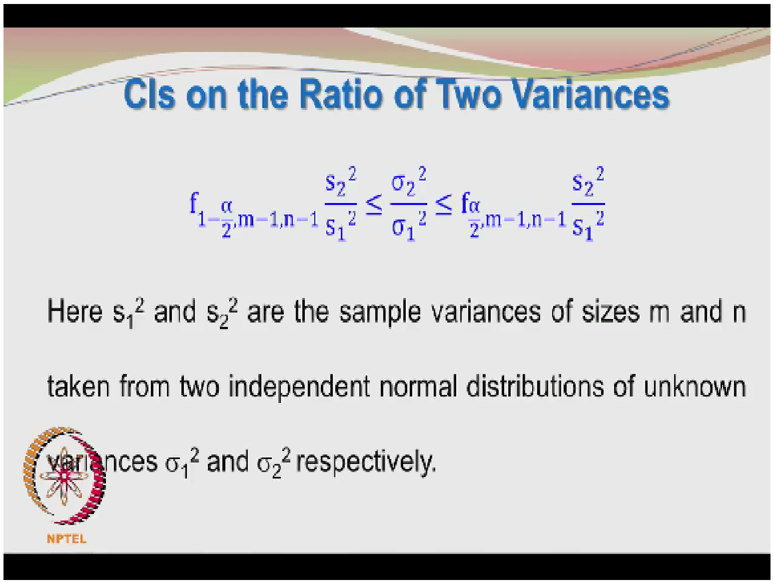
The 100*1-alpha percentage confidence interval on the ratio of the variances sigma2 squared/sigma1 squared=f of 1-alpha/2 m-1 n-1 S2 squared/S1 squared <= sigma2 squared/sigma1 squared <= f of alpha/2 m-1 n-1 S2 squared/S1 squared. So this is the confidence interval for the ratio of the 2 variances, sigma2 squared and sigma1 squared.

The important thing to note here is whether we are using sigma1/sigma2 whole squared or sigma2/sigma1 whole squared and depending upon which is present, this also will get affected. So you had to be a bit careful whenever you are setting up the confidence interval for the ratios of 2 variances. You have to make sure that the correct numerator and denominator degrees of freedom are present in the upper and lower limits.

If you look at this, so if I simplify this sigma2 squared will go to the numerator, sigma1 squared with come in the denominator and this will remain as they are. Then I can take multiply S2 by S1 whole squared throughout and so this will cancel out. So you will have f of 1-alpha/2m-1 n-1*S2 squared by S1 squared <= sigma2 squared/sigma1 squared <= f of alpha/2 m-1 n-1 and S2 squared/S1 squared that is = 1-alpha.

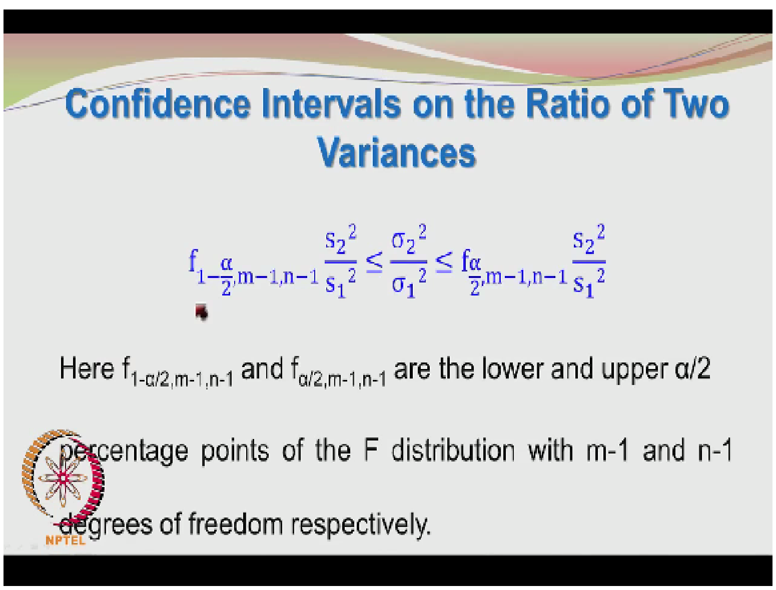
So the confidence interval is defined after the sample is taken in the S1 squared and S2 squared are known in the fashion shown here. So once the 2 samples have been taken, we can compute their sample means and the sample variances. So we can compute these 2 values and then we can define the confidence interval around sigma2 squared/sigma1 squared.

(Refer Slide Time: 28:26)



Here S1 squared and S2 squared are the sample variances of size m and n taken from 2 independent normal distributions of unknown variances sigma1 squared and sigma2 squared. So again we have to assume that the parent populations are normal.

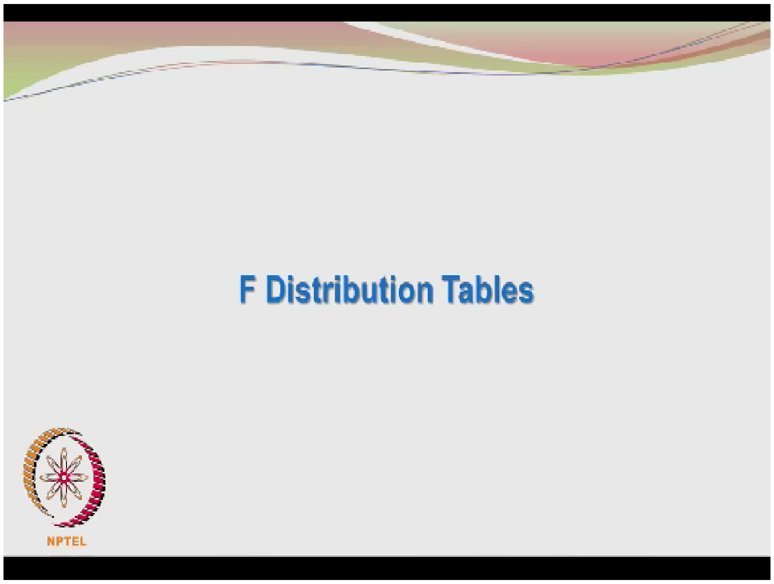
(Refer Slide Time: 28:51)



Here we may define f of 1-alpha/2 m-1 n-1 and f of alpha/2 m-1 n-1 as the lower and upper alpha/2 percentage points of the F distribution with m1 and n-1 degrees of freedom respectively. So here we are talking about the lower tail probability corresponding to f of 1-alpha/2. We are finding out the alpha such that S2 squared/S1 squared represents a number and the probability below this value is alpha/2. So that is why it is called as the lower tail.

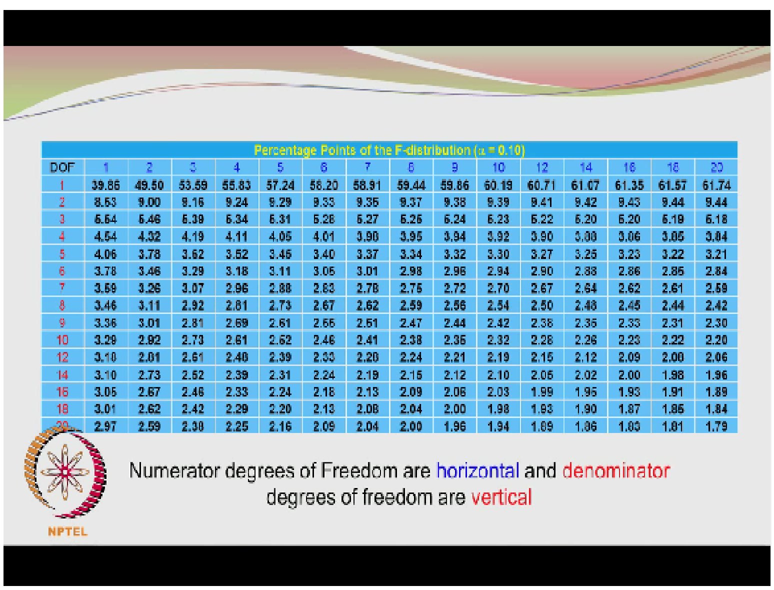
So in the region below this number, the area under the curve is alpha/2 and the upper tail probability would obviously be 1-alpha/2. So we are identifying 2 numbers f of 1-alpha/2 m-1 n-1, f of alpha/2 m-1 n-1 alpha/2 percentage points of the F distribution with the m-1 and n-1 degrees of freedom respectively.

(Refer Slide Time: 30:22)



So now let us take a quick look at the F distribution tables.

(Refer Slide Time: 20:29)



Percentage Points of the F-distribution (α = 0.10)

DOF	1	2	3	4	5	6	7	8	9	10	12	14	16	18	20
1	39.86	49.50	53.59	55.83	57.24	58.20	58.91	59.44	59.86	60.19	60.71	61.07	61.35	61.57	61.74
2	8.53	9.00	9.16	9.24	9.29	9.33	9.35	9.37	9.38	9.39	9.41	9.42	9.43	9.44	9.44
3	5.54	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.24	5.23	5.22	5.20	5.20	5.19	5.18
4	4.54	4.32	4.19	4.11	4.05	4.01	3.98	3.95	3.94	3.92	3.90	3.88	3.86	3.85	3.84
5	4.06	3.78	3.62	3.52	3.45	3.40	3.37	3.34	3.32	3.30	3.27	3.25	3.23	3.22	3.21
6	3.78	3.46	3.29	3.18	3.11	3.05	3.01	2.98	2.96	2.94	2.90	2.88	2.86	2.85	2.84
7	3.59	3.26	3.07	2.96	2.88	2.83	2.78	2.75	2.72	2.70	2.67	2.64	2.62	2.61	2.59
8	3.46	3.11	2.92	2.81	2.73	2.67	2.62	2.59	2.56	2.54	2.50	2.48	2.45	2.44	2.42
9	3.36	3.01	2.81	2.69	2.61	2.55	2.51	2.47	2.44	2.42	2.38	2.35	2.33	2.31	2.30
10	3.29	2.92	2.73	2.61	2.52	2.46	2.41	2.38	2.35	2.32	2.28	2.26	2.23	2.22	2.20
12	3.18	2.81	2.61	2.48	2.39	2.33	2.28	2.24	2.21	2.19	2.15	2.12	2.09	2.08	2.06
14	3.10	2.73	2.52	2.39	2.31	2.24	2.19	2.15	2.12	2.10	2.05	2.02	2.00	1.98	1.96
16	3.05	2.67	2.46	2.33	2.24	2.18	2.13	2.09	2.06	2.03	1.99	1.95	1.93	1.91	1.89
18	3.01	2.62	2.42	2.29	2.20	2.13	2.08	2.04	2.00	1.98	1.93	1.90	1.87	1.85	1.84
20	2.97	2.59	2.38	2.25	2.16	2.09	2.04	2.00	1.96	1.94	1.89	1.86	1.83	1.81	1.79

So a colourful table is presented before you. Please read the text at the bottom, numerator degrees of freedom or horizontal and the denominator degrees of freedom are vertical and they are shown in red. The numerator degrees of freedom go horizontally in this direction and they are shown in blue. Again here what we have to do is look at the parameter given here, alpha = 0.1.

So for an alpha value of 0.1 that means the probability value is given as 0.1. More specifically, the upper tail probability value is given as 0.1. Then if the numerator degree of freedom is 4 and the denominator degrees of freedom is 5, so I will go like this. I will go like this. So the required F value is 3.52. So probability of the F random variable > 3.52 will be = 0.1. The next table gives an alpha value of 0.05. Earlier we saw the alpha value of 0.1.

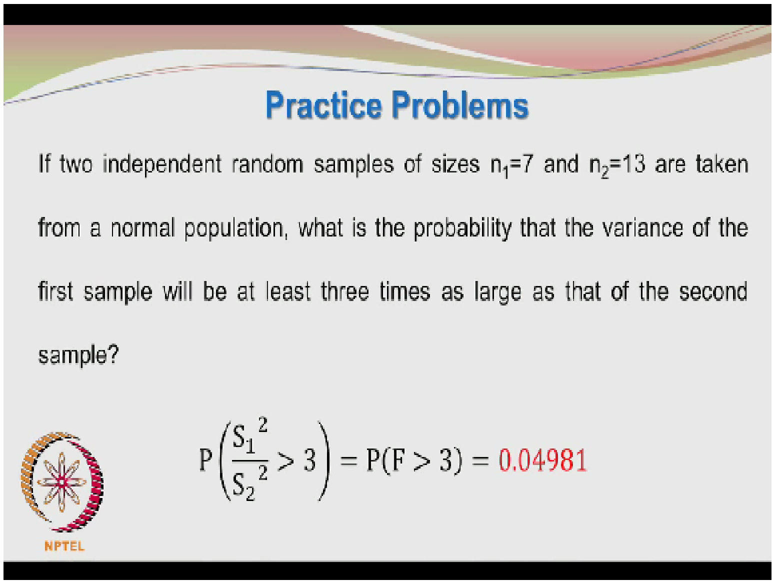
Now it is alpha = 0.05. So we are having a more and more information that has to be presented. Earlier we were a looking at the normal distribution chart, the standard normal distribution chart and we had only 1 chart or 1 table and there you are given the EZ values and you could read out the probability but when you went for the P distribution, you had the additional parameter, the degrees of freedom.

You had one additional parameter that is the degrees of freedom and then what you had to do was, identify a probability value, identify a degree of freedom and then find the corresponding P value. Same thing happened in the chi-square distribution but when you look at the F distribution, you have the numerator degrees of freedom and denominator degrees of freedom and so you need more tables or more charts. So one chart is for one probability value.

So here you have alpha = 0.05. Numerator degrees of freedom, denominator degrees of freedom. So once you know the alpha value or the probability value and you know the numerator degrees of freedom, one of these and the denominator degrees of freedom, you can then find the corresponding F value. What is the F value that will lead to a probability of 0.05 with 6 and 3 degrees of freedom?

So I will locate 6 which is the first degree of freedom located in the numerator and then I will locate 3, come here and I will get 8.94. Similarly, other F values for other degrees of freedom maybe estimated. Similarly, we have another table where the probability is different. Now the alpha = 0.01. What is the F value which will give me a probability of 0.01 if my degrees of freedom are 7 and 8. So 7 numerator, 7 and then 8 is here, so there corresponding F value is 6.18.

(Refer Slide Time: 35:20)

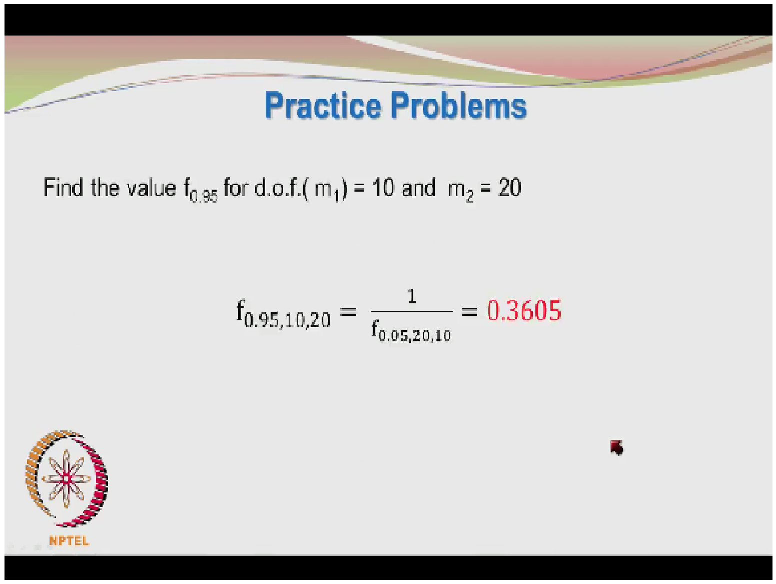


So now let us look at a few problems. If there are 2 independent random samples of sizes n1=7 and n2=13 and they are from a normal population, from a single normal population, what is the probability that the variance of the first sample will be at least 3 times as large as that of the second sample? So we define the F random variable as S1 squared/sigma1 squared/S2 squared/Sigma1 squared, okay because we are talking about the same population.

So the same Sigma1 squared will be in the numerator as well as in the denominator and hence will get cancelled out and we are only left with S1 squared/S2 squared. What is the probability that this ratio will be > 3? So probability of F>3 with 7 numerator degrees of freedom that is incorrect, with 7-1 that is 6 numerator degrees of freedom and 13-1=12 denominator degrees of freedom can be shown to be 0.05.

Coming again probability of S1 squared/S2 squared>3=probability of the F random variable with 6 numerator degrees of freedom and 12 denominator degrees of freedom = 0.04981 or more or less 0.05.

(Refer Slide Time: 37:12)



Find the value of F0.95 for m1=10 and m2=20. In the previous case, we were having the sample sizes. So we had to find m1 as 6 and M2 as 12. Now we are directly given the degrees of freedom themselves that is 10 and 20 for the numerator degree of freedom and the denominator degree of freedom. So F of 0.95,10,20=1/f of 0.05,20,10 and f of 0.05,20,10 maybe found from the table. Let us go to that, f of 0.05,20,10. So 0.05, 20 is numerator, denominators is, let me just check the degrees of freedom, 20,10, 0.05 and then we have 20, denominator is 10.

So you find out the value as 2.77 and then you put 1/2.77, approximately 1/3 0.33 since 1/2.77, it will be slightly higher and that the value is 0.3605. So at present, we have completed the brief discussion on the F distribution. We have done the T distribution, the chi-square distribution and the F distribution. What we can now do is look at a few example problems and see how these distributions may be applied. So we will take a small break here and then we will continue with the example problems.