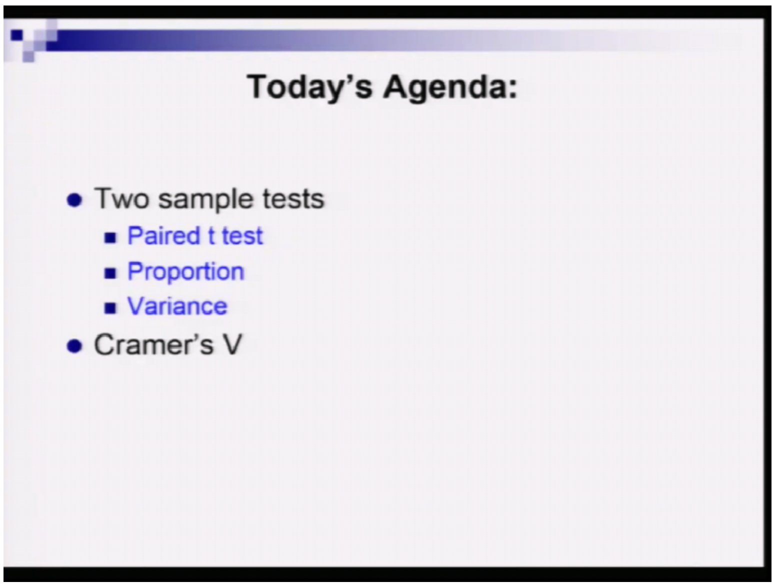
Applied Statistics and Econometrics
Professor Deep Mukherjee
Department of Economic Sciences
Indian Institute of Technology, Kanpur
Lecture-14
Hypothesis Testing (Part IV)

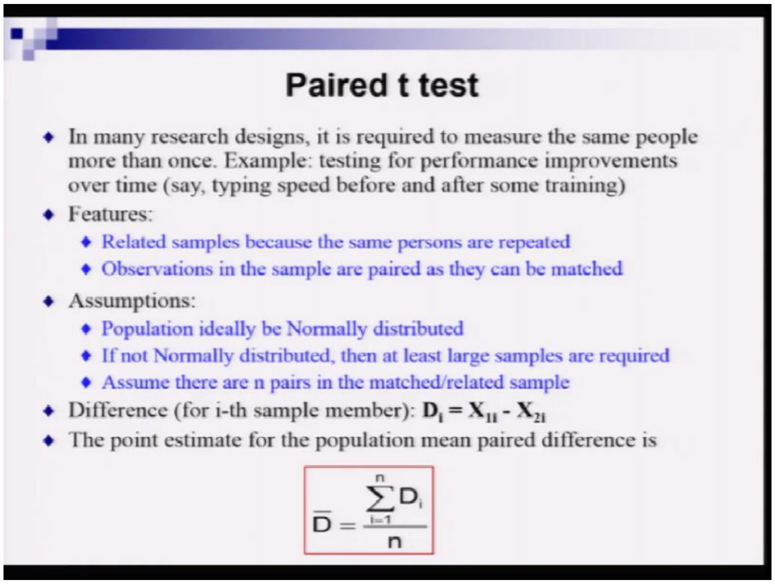
Hello, friends. Welcome back to the lecture series on Applied Statistics and Econometrics. So today, we are going to continue our discussion on hypothesis testing and we are going to look at a couple of new tests. So, let us have a look at today's agenda item.

(Refer Slide Time: 00:31)



So, in today's lecture, we are going to talk about a very useful test called paired t test. Then we are also going to look at the other variations of 2 sample tests, namely proportion and variance tests. And finally, we will end our discussion by briefly discussing a very interesting concept called Cramer's V.

(Refer Slide Time: 01:38)



So, our first topic is paired t test. So, what does it do actually? If you want to describe it in simple language, so, sometimes we study the same group of subjects or elements in repeated samples and we may be interested to see how the mean is changing from one sample to the other. So here, as elements in one sample is matched with the element in another sample that is why it is called the paired sample or matched sample and we are going to now describe the t test, which could deal with these type of samples.

So, in many research designs, it is required to measure the same people more than once. So, here is an example, suppose manager of an office or your boss, he has decided to send some staffs for training so, that they can improve their typing speed. So, before he sent his staff to the training, the typing speed was recorded and then, after the training, when the person rejoins the job again the typing speed is recorded. So, if you send some 10, 15 or 20 staff for the training program, so, there is a before training mean typing speed and then there is an after training typing speed. So, you may be interested to see whether there is an improvement in the overall mean typing speed for this group of staff.

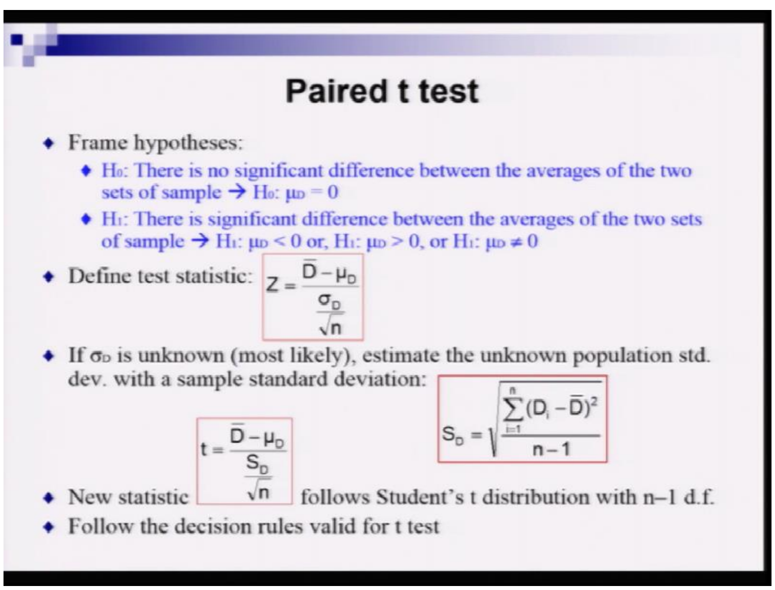
So, here, you have to apply paired t test and now I am going to briefly discuss about some features and assumptions. So, I have already mentioned these features that observations in the sample are appeared as they can be matched across the sample. So that we know we have discussed.

Now, let me state the assumptions clearly. So, we assume that population ideally be normally distributed, if it is not normally distributed, then at least large samples are required. And then, you assume that there are n pairs in the matched or related sample. So, it is better that you have n greater than or equal to 30.

So, to conduct the test, the first step is to compute the difference and this is for the i-th sample member. So here, I define this difference as di as x1i minus x2i. So x1 and x2 are basically 2 different groups. So, suppose this typing speed is a variable. So, 1 is basically after the training and 2 is before the training or you can think of several other examples,. But i is basically common, so, that is basically i-th sample member who is common in these 2 groups x1 and x2.

So then, how do I define my point estimate for the population mean pair difference? So here is the formula. So, D the bar is the notation that I am introducing here as the point estimate and that is basically nothing you have to take the arithmetic mean of that variable di.

(Refer Slide Time: 04:21)



Now in the next step, you have to frame hypothesis as usual. So, here I start with the null hypothesis that there is no significant difference between the averages of that 2 sets of sample. And then, I say that okay, my h naught can be represented as mu d equal to 0. So, d is basically the difference variable that I have created and mu is basically the population mean.

Now, alternative hypothesis can be set in 3 different ways. So here, we define the alternative hypothesis as there is significant difference between the averages of the 2 sets of sample. And there can be 3 types of mathematical form that 1 can issue. So first 1 is going to give you a left-tail test, where you write alternative hypothesis as mu d less than 0. And the second type could be the right-tail test. And there you write your alternative hypothesis as mu d is positive. Or you can have a two-tailed test where you can write h1 as mu d naught equal to 0.

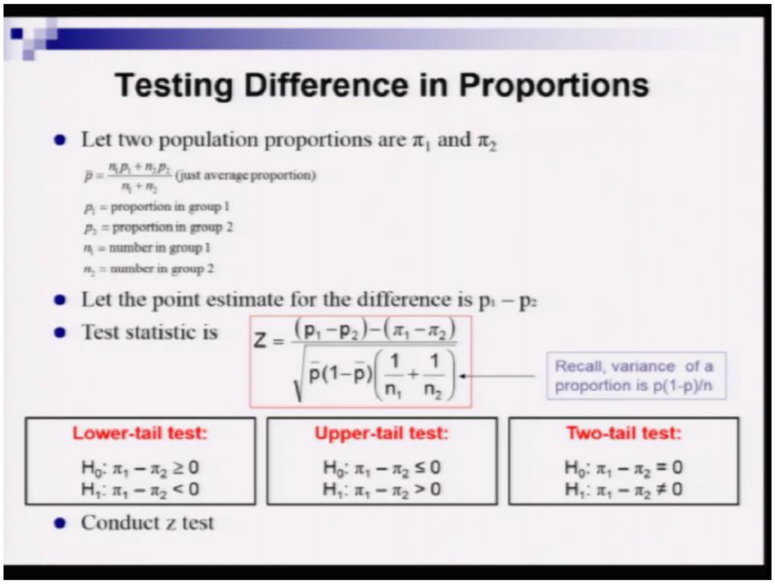
So now in this case, you have to define the test statistic as the next step and here is the test statistic that is basically a z score. And then, you have d bar that is the sample mean of the difference variable d. And then you have to take the difference between the sample mean of d and the population mean mu d. So, if you work with the above shown null hypothesis then mu d will of course take value 0, when you compute the test statistic.

And this difference now shall be divided by the sigma d divided by root n. So, the sigma d is of course, the unknown population standard deviation and most likely it will remain unknown. So, then what to do? So, you can actually get a proxy for sigma the. And the proxy could be defined as the sample standard deviation, which is defined it sd. And I am showing you the formula in the red box. And that is not uncommon to you, you have seen how 1 can compute the sample variance. So, we are using the same formula, only in the place of x is now we are writing d.

So, once that sd is computed, in the next step you write the expression for the new statistic and that is a t statistic. And I am showing you the expression in the box and this will follow a student’s t distribution with n minus degrees of freedom. And then, you follow the decision rules valid for the t test that we have discussed before. It may also be interesting to go for a hypothesis testing for population proportion. So, sometimes it is interesting to compare the population proportions that come from 2 different populations.

So, you may get a sample from population 1 and you may get another sample from population 2. And you look at the sample proportion, they look different to you, but are they statistically different? That could be an interesting research question. So, in the next hypothesis testing, we are going to discuss this particular issue.

(Refer Slide Time: 07:50)



So, we begin by introducing some notations. So, let 2 population proportions are pi1 and pi2. And then, I define something as p bar, that is a new thing that we are defining in this course and that is basically a weighted average of the sample proportions. So, here, if I am dealing with 2 groups, group 1 and group 2, the sample proportions are p1 and p2. So, what is p1 and what is p2? So, if there are x1 number of elements satisfying some criteria out of n1 sample in group 1, then p1 is defined as x1 divided by n1. And similarly, 1 can define p2 as well.

So, now, as there are n1 and n2 number of observations into different groups, respectively, you take weighted average and then you get the overall sample proportion p bar. And then, once that is calculated, you now define your point estimate for the difference and now, you define that to be p1 minus p2 and next you decide on the test statistic and then you have the complicated z formula.

Well, it may look complicated initially, but it is not. Because if you recall certain things from the previous lectures, it is very common thing that we are still following here even in this particular slide. So, when you actually write a test statistic, you have to, in the numerator you have to first you talk about the difference between the sample statistic and the unknown population parameter value.

So, here p1 minus p2 is basically that point estimate that is basically as a statistic you can assume. And then, pi1 minus pi2 is basically the unknown difference between population proportions, pi1 and pi2. So, the same philosophies or the concept is applied here in the numerator. And in the denominator, what are we seeing? So, if you remember, the variance of a proportion is basically p times 1 minus p divided by n, that we have seen earlier. So, now p is unknown. So, we have to get a proxy measure for p from the sample and that is basically your p bar, the overall sample proportion.

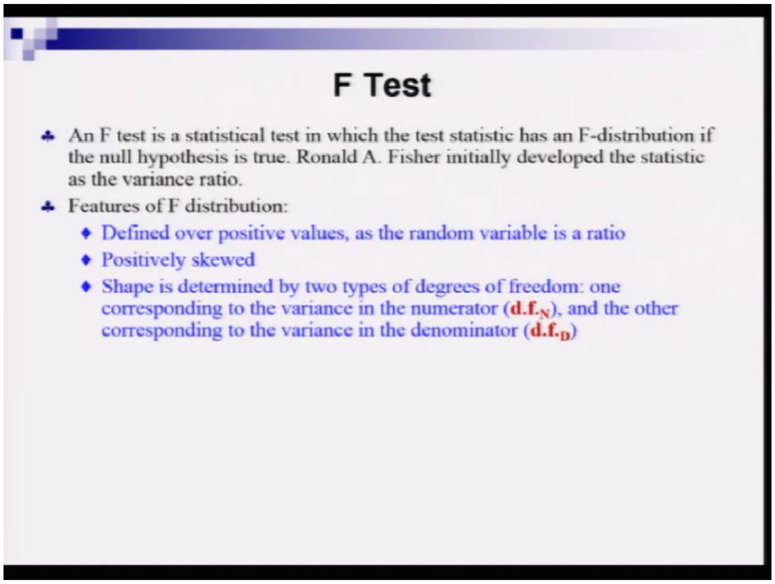
So basically, now you have 2 different sample sizes n1 and n2. So, you see the term that you are seeing under the square root that is basically the combined variance of 2 different samples. So of course, here also, we can see that there could be 3 different scenarios, 1 is lower-tail or left-tail test; 1 is upper or right-tail test; and there could be 1 two-tailed test.

So, here first we are going to talk about the lower-tailed test. And I am showing you, how you can write the null and alternative hypothesis. So, nothing new here. Similarly, 1 can frame the null and alternative hypothesis for the upper-tail test, nothing new here as well. And then finally, you have that two-tailed test.

And you have to first decide which particular test you want to conduct. So, accordingly, you frame your hypothesis. And then, after that, you have to conduct a standard z test, because here, your test statistic is z. And as you are assuming that you have large samples so, n1 and n2 ideally is greater than or equal to 30. So, you can actually conduct as a z test.

So now, we are going to discuss the case of F test, we have not seen F test before. So, we are going to introduce this concept for the first time. And you will see later on in the course we will find very nice applications of d test. So, in the next slide, I am going to show you the fundamental steps of conducting an F test. And I am not going to get into details of statistics regarding the F test, I am just going to talk about the most important things.

(Refer Slide Time: 12:20)

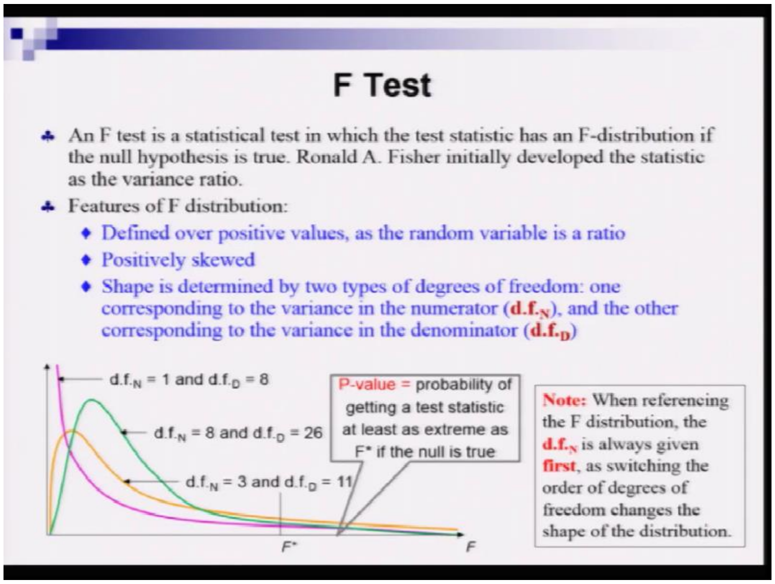


So, let me first introduce the F test in the most simple possible language. So it is test for which the test statistic follows an F-distribution, if the null hypothesis is true and famous statistician, Ronald A. Fisher actually first developed the statistic as the variance ratio. If you remember, the way I defined the F value and the F distribution, while I was talking about the children of the normal distribution, then it was the ratio of the variances and that is, that was the form that Ronald Fisher introduced. And in his respect, this distribution finally was named F distribution.

So, there is another interesting feature that in most cases, the F distribution is a positively skewed distribution. But the shape of the F distribution is going to change because F is a very complicated distribution, it has 2 degrees of freedom.

Remember, in the chi-square case, we had 1 degrees of freedom. Now here, we have 2 types of degrees of freedom, because it is a ratio random variable. So, for the numerator, there is 1 degrees of freedom and for the denominator there is another degrees of freedom.

(Refer Slide Time: 13:39)



So, the shape is determined by d.f.N, which is the degrees of freedom corresponding to the variance in the numerator and the d.f.D, which is corresponding to the variance in the denominator of the ratio of 2 variances.

So, at the bottom part of the slide, I am going to show you a small diagram, a simple diagram where I am showing you various cases of F PDF. So, here you see the pink one which is looking almost like a rectangular hyperbola is F PDF for d.f.N equal to 1 and d.f.D equal to 8.

And you see if I now increase the d.f.N and d.f.D values to 3 and 11 respectively, then I get these orange colored PDF shape, which is a typical positively skewed distribution and if I continue to increase the d.f.N and d.f.D values to 8 and 26, then I get a PDF of higher height and it has got higher density in the right-tail also and that is given by the green colored PDF curve.

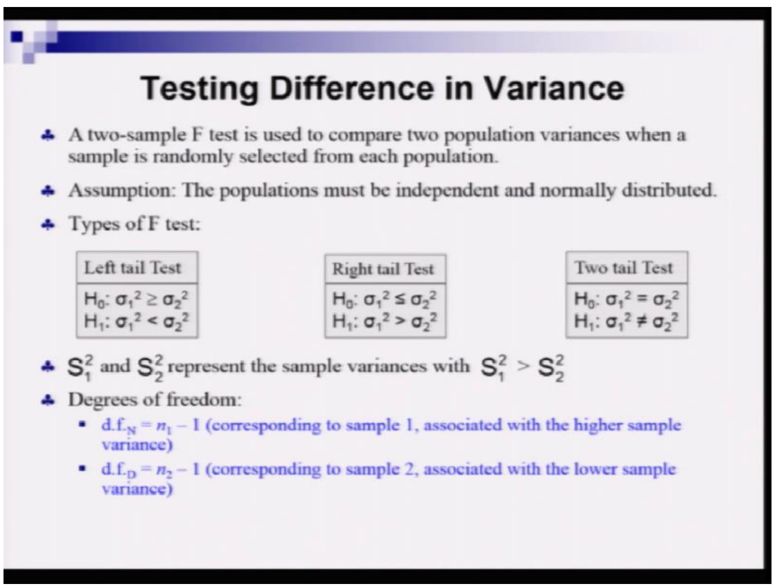
Now, I am going to tell you here about one statutory warning or you can see that as a cautionary note. So, when you are referencing F distribution, when you are trying to get the critical values from an F table, then d.f.N or the degrees of freedom corresponding to the numerator is always given first. So, you have to look at the column first and once you fix the column for the numerator, degrees of freedom d.f.N, then you have to come to the row for d.f.D, which is basically the degrees of freedom for the denominator.

And this is very important. Why? Because switching the order of these d.f.N and d.f.D will have a big impact on the shape of the distribution and the corresponding critical values and probability values.

Now, in this context of F distribution, let us look at the very important concept of p value, once again. So here, now, let us concentrate on this diagram again. And now, we are going to talk about a specific value of F, which is the calculated value of test statistic. And let me denote by F star. So, now you erect a vertical line or draw a vertical line on F star. And then of course, it will partition your probability density function into 2 parts, 1 in the right hand side and 1 on the left hand side.

So now, when you concentrate on the area of right hand side, right with respect to F star, then basically that area below the F PDF curve, gives the p value. So, how do I interpret these p value? So here in this context, it is the probability of getting test statistic at least as extreme as F star if the null hypothesis is true.

(Refer Slide Time: 17:30)



Now, we are going to look at 1 example of F test. And that is basically when we want to compare 2 population variances, we do not know the true values of the population variances. But we want to draw some inference based on the sample variances. So how do you conduct a statistical testing to solve this problem?

So here are 2 sample F test is used to compare 2 population variances, when a sample is randomly drawn from each population. Here, we have to make a very strong assumption that the populations must be independent and normally distributed. And if you remember, our previous lectures, the F test could be of 3 types here, 1 of the 3 types, and they are namely left-tail test, right-tail test and two-tailed test. And I am showing you the null and alternative hypothesis under both categories, so you see, these expressions are given in these 3 boxes.

So, the left hand box talks about the left-tailed test and you see sigma 1, 2, these are the notations I am using to denote the population standard deviations for population 1 and population 2. And if you follow that notation, then of course, these expressions are easy to follow. And then, in the center, I have right-tail test description of null and alternative hypothesis. And finally, at the right-hand side, the third box gives me the null and alternative hypothesis for a two-tailed test.

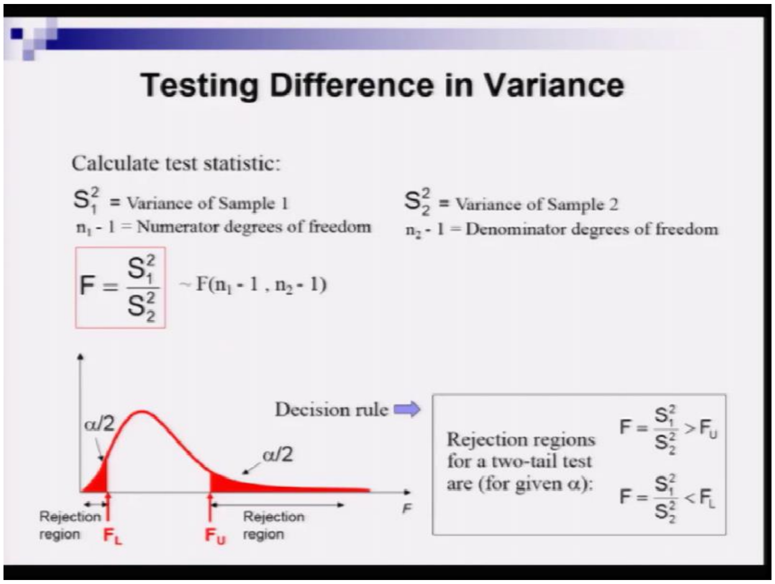
Now, the question is that we do not have any a priori knowledge about the population variance values. So, we do not know sigma 1 square and sigma 2 square. So then, how to proceed? We have to draw a sample and we have to calculate the sample variances from these 2 samples and let me call them s1 square and s2 square.

So, after calculating s1 square and s2 square, the sample variances most likely we are going to see that some, 1 of this is higher than the other. So, you call that higher variance corresponds to the population 1 so you name that population as population 1 for which you observe the highest value or the higher value of sample variance. And so, we can write s1 square is greater than s2 square. This is important to conduct an F test for checking the difference in variances.

Now, it is time to look at the degrees of freedom issue. So, if you remember that when we are conducting F test through F distribution, it has 2 degrees of freedom, 1 is associated with the numerator and other 1 is associated with the denominator. So, let us have a look at them, what they are? So here d.f.N is n1 minus 1. So that is basically the number of observations in sample 1 minus 1 and d.f.D, that is basically corresponding to sample 2 and it is defined as number of observations in sample 2 minus 1.

So, I am not going to give you sentences where I am typing what to do next what to do next. So, here I am going to take a graphical approach to explain how to conduct an F test. We already have spoken about that diagram, 2 slides before. So, you will see a practical application of that slide again here in this slide, where I am going to describe this F test procedure again

(Refer Slide Time: 20:54)



So, this slide will start by calculating the test statistic value. So, we have to define the test statistic first. So, the test statistic is defined as capital F and that is the ratio of the sample variance from population 1 and sample variance of population 2. So, remember here that population 1 is that population, where the sample variance is higher when you compare 2 sample variances.

So, that is what we have written and these particular test statistic follows an F distribution with 2 degrees of freedom, namely in n1 minus 1 and n2 minus 1. So, I have actually explained them the previous slide. So next, you have to set the level of significance at alpha and then, you need to set the decision rule. So, for a particular level of alpha, as the degrees of freedom are given for both numerator and denominator, it is not a difficult task to get the critical values, FU and FL. So, FU is basically the critical value for the right-tail or it can be called critical value of upper-tail. And correspondingly we can find FL that is basically the critical value for left-tail or the lower-tail and that is denoted as FL.

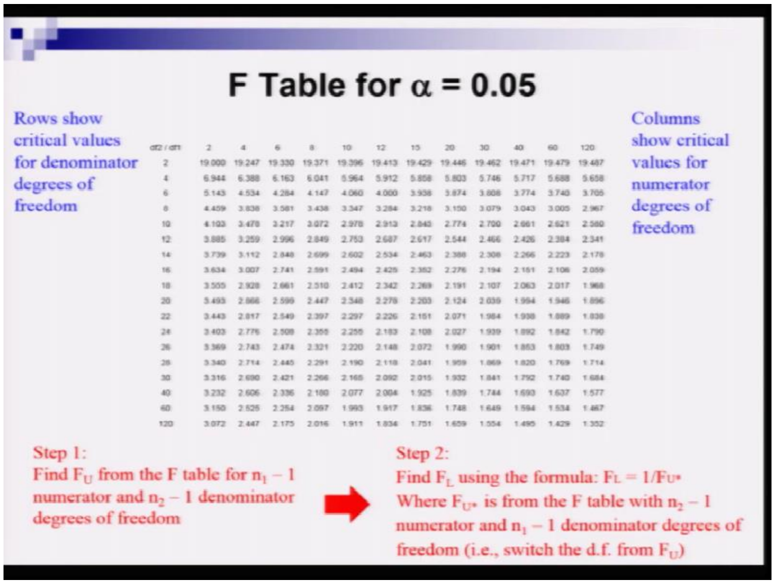
So, now, I set a decision rule that if the calculated value of F is greater than FU then you can reject your null hypothesis or if the calculated value of F is less than the critical value FL, then also you can reject your null hypothesis. So, that is basically shown here in this diagram at the left-hand side and the bottom part of the slide.

So here, it is basically the same diagram that I have shown you, maybe a couple of minutes back. But here, I am indicating the rejection regions again, corresponding to or relative to FL and FU, the critical values and then, the task at hand is to get the FL and FU, in that case, you will be able to take a decision for this sample test.

Now, finding F values from statistical tables is a very challenging task, why? Because for a particular level of alpha, you have 2 degrees of freedom, so when you think about that matrix form of the F values, so in row you will have 1 set of degrees of freedom values, and on the columns you will have other set of degrees of freedom values. But note that, that matrix structure or the table will be valid for 1 particular alpha value. If you change alpha, then again, the same table will be reproduced but with different numbers in the sales as you have changed the alpha value.

So, F table it is very difficult to give you the entire picture in 1 slide, I am going to show you a glimpse of 1 particular F table for 1 particular level of significance. But there are many textbooks where you can find the longer versions of F table or you can use software or online calculators to find the critical values or the probability score in the F table. So, let me now take you to a short glimpse of a particular F table.

(Refer Slide Time: 24:48)



So here, I set alpha value add 0.05. And here, the rows show the critical values of the denominator, degrees of freedom. So, d.f.D and the columns show the critical values for the numerator of degrees of freedom. So, it is basically d.f.N. So, as I told you, 2, 3 slides before that you have to first look at the column and then, you have to come down rows and then you have to refer to the denominator degrees of freedom. So, that is the way you have to read the table.

But this is the table for right-tailed test. So, if you now have to conduct the two-tailed test, how do you proceed? it is actually not very simple. So, I am trying to express this in words. So now, look at the bottom of the slide, here I have written 2 steps. So, in step 2, you have to first identify a FU from the F table for n1 minus 1 numerator and n2 minus 1 denominator degrees of freedom.

And then in step 2, you have to find the FL, the critical value at the left-tail by using a formula and that is given by FL equal to inverse of FU star. Now, this FU star is not the FU that you have computed in step 1, this is somewhat different. So, what is it? So here, FU star actually is to be found from the F table with n2 minus 1 numerator and n1 minus 1 denominator degrees of freedom. So, you have to basically switch the degrees of freedom from the calculation of FU.

So, we are going to end today's discussion, I thought to cover Cramér's V in today's lecture. But I had a second thought and I think that I will cover that concept in the lecture on the relationship between variables. So, wait for the next lecture. Thank you.