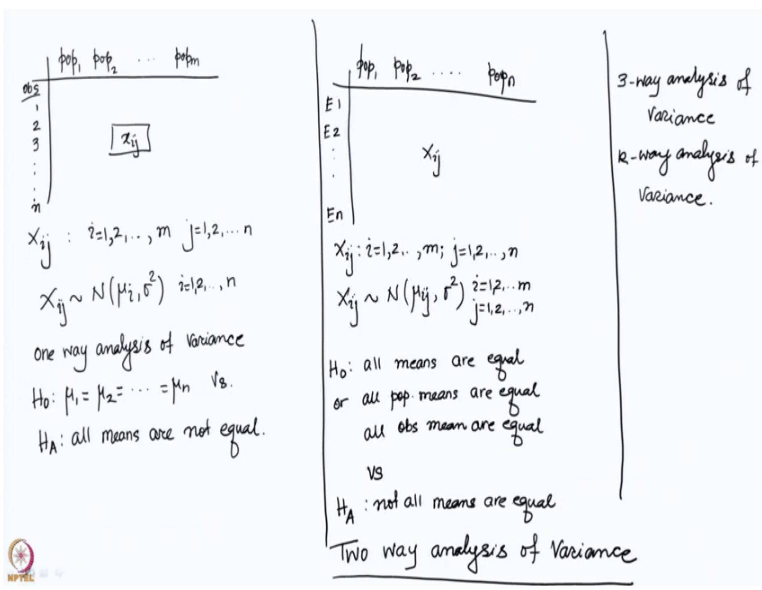
Dealing with Materials Data: Collection, Analysis and Interpretation
Professor M P Gururajan,
Professor. Hina A Gokhale.
Department of Metallurgical Engineering and Materials Science
Indian Institute of Technology, Bombay
Lecture 86
Analysis of Variance – II

Hello and welcome to the Dealing with Materials Data course. We are going through the sessions on Analysis of Variance. The first session what we looked into is called the One Way Analysis of Variance.

(Refer Slide Time: 0:40)



Let us review it quickly, we had a matrix, we had m population, population 1, population 2, dot, dot, dot, population m. And under each population we had several observations. 1, 2, 3, etc, n. So, we had n observation under m population and a typical data value we call it Xij, right? Where we say that Xij is an observation with i running from 1, 2, 3, up to m, that is m population and j represents the j, n observations. And now, we assume that Xij is distributed as normal with mean Mui and variance sigma square, again for i is equal to 1 to n.

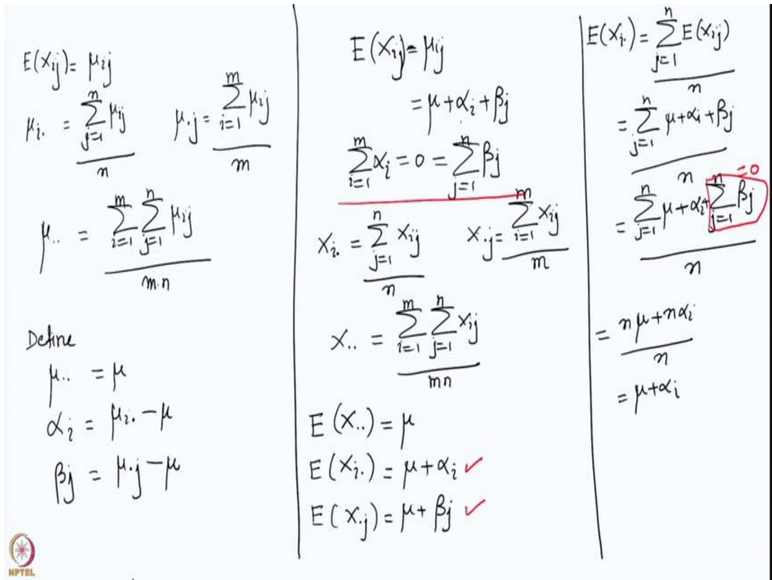
In other words, we said that, the distribution of this typical element Xij depends only on the population and not on the observation. Therefore, it is called One Way Analysis of Variance. And our hypothesis of interest was Mu1 is equal to Mu2 is equal to dot, dot is equal to Mun. All Mus are same versus the alternate hypothesis was that all means are not same.

And now, we what we want to consider is suppose we have a population, same population 1, population 2, etc etc population n. And we have some observation which I called observation 1, observation, 2 and observation n and a typical value Xij is there. Again Xij is such that i varies from 1, 2, etc to m and j varies from 1, 2, etc to n.

But the difference is, I am going to now assume that Xij is distributed as normal with mean Mu ij, sigma square for i is equal to 1, 2, 3 n and j is equal, sorry m and j is equal to 1, 2, 3 n and we want to have H0 as all means are equal or we can have another one which says that all population means are equal or it could be all observation means are equal versus the alternate that not all means are equal.

Bit complicated, is not it? So it is a kind of a we are now generalizing it. This case is called Two Way Analysis of Variance. This is called Two Way Analysis of Variance. Just for your information, today we are going to tackle how to solve or how to do this analysis, but then the way we are going to do it, one can also work out Three Way Analysis of Variance or in very general terms, k minus way Analysis of Variance.

(Refer Slide Time: 6:52)



So, that is going to be our today's plan. Let us start, so we have a population and we are assuming that expected value of Xij is equal to Mu ij, and now we have to simplify this, because our hypothesis is quite complicated, if we want to check that all population means are equal or all observation means are equal, it is kind of difficult. So we need to simplify this expression of Mu ij.

So we start it this way, we say Mui dot is equal to summation i is equal to 1 to m, Mu, sorry j is equal to 1 to m Mu ij divided by n. This should be, let us correct ourselves, this should be n. This is Mui dot, similarly you can have Mu dot j, whatever we are averaging on, we are putting a dot in that place, so it is summation of i is equal to 1 to m, Mu ij divided by m. And finally Mu dot dot is equal to summation i is equal to 1 to m, j is equal to 1 to n, Mu ij divided by m times n.

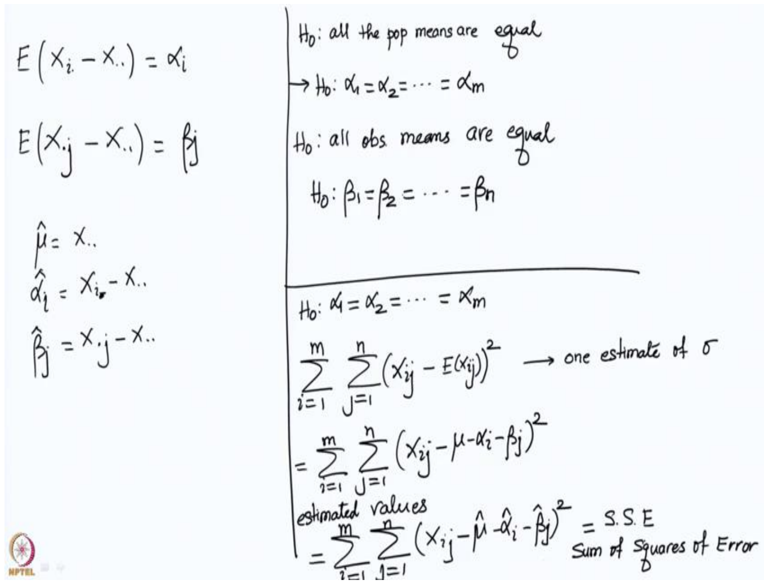
Now, you define, Mu dot dot is equal to Mu, alpha i is equal to Mu i dot and beta j is equal to Mu dot j. Then, please see that expected value of Xij is now Mu ij which can be written as Mu plus plus alpha i plus plus beta j. We have sort of divided out, sorry I have made a mistake, this minus Mu and this minus Mu, yes. So now we have sort of divided out the complete mean Mu ij into ith component, jth component and a common component Mu.

And also notice that now summation of alpha i, i is equal to 1 to m is equal to 0, it is also the case with summation j is equal to 1 to n beta j. So this is the condition and now, also let us define the corresponding value of Xi, so I called Xi dot is equal to summation of Xij, j is equal to 1 to m divided by m, Xj dot, sorry, X dot j, let me write it correct. X dot j is equal to summation i is equal to 1 to m, this is a mistake, let me correct it. It should be summation j is equal to 1 to n, divided by n and here it will be Xij divided by m population.

And then again X dot dot will be summation i is equal to 1 to m, summation j is equal to 1 to n, Xij divided by m n. Then, you can see that expected value of X dot dot is Mu, expected value of Xi dot is equal to Mu plus plus alpha i and expected value of X dot j is equal to Mu plus plus beta j. How will it be? Because it will come, let us change the ink color, because of this condition, these two equations can be derived.

Let us show it in one case, let us try to find, sorry. Let us try to find expected value of Xi dot which is, summation Xi dot is j is equal to 1 to n, expected value of Xij divided by n and if you take summation of that it is summation of j is equal to 1 to n, this is Mu plus plus alpha i plus plus beta j and divided by n and therefore, it will become summation of j is equal to 1 to n, Mu plus plus alpha i, remember that summation plus plus summation j is equal to 1 to n beta j divided by n, and note that this quantity is equal to 0 and therefore what we have is, this is equal to n times Mu plus plus n times alpha i divided by n which Mu plus plus alpha i and that is what is written here, so this is how it has been derived.

(Refer Slide Time: 14:52)



So, these are our notations, so from there we can work out that expected value of Xi dot minus X dot dot is alpha i and expected value of X dot j minus X dot dot is beta j, I think this calculation can be worked out and therefore, we can define Mu hat is equal to X dot dot alpha i hat is equal to Xi dot minus X dot dot and beta j hat, these are the estimators for them, so this is X dot j minus X dot dot.

Now, let us try to rewrite the hypothesis that we wish to test. So hypothesis that we wish to test for example which says that, all the population means are same. This translates into hypothesis that alpha 1 is equal to alpha 2 is equal to alpha m. Please notice that we have made all these transformations, mathematical transformation only to simplify the hypothesis.

So if we say that, hypothesis is that, all observation means are equal, then our H0 is actually beta 1 is equal to beta 2 is equal to beta n. So, this is how we translate it. Let us take one of the hypothesis, we take let us want to test hypothesis that alpha 1 is equal to alpha 2 is equal to so on alpha m. Now you see, again we will have two estimates for sigma, remember that sigma square, the population variance is common for all the observations.

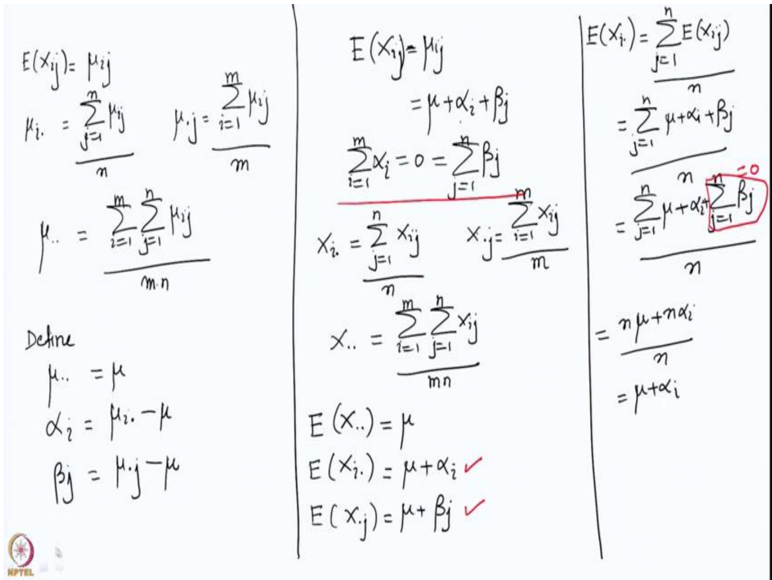
So, we have now two estimates of sigma. One estimate is summation i is equal to 1 to m, j is equal to 1 to n Xij minus expected value of Xij whole square. This is one estimate of sigma. This with an appropriate divider, this will give you one estimate of sigma. This can be written as summation i is equal to 1 to m, summation j is equal to 1 to n, Xij minus Mu minus alpha i minus beta j whole square and if you replace with the estimated values, with estimated values, this is going to be summation i is equal to 1 to m, summation j is equal to 1 to n, Xij minus Mu hat minus alpha i hat minus beta j hat whole square.

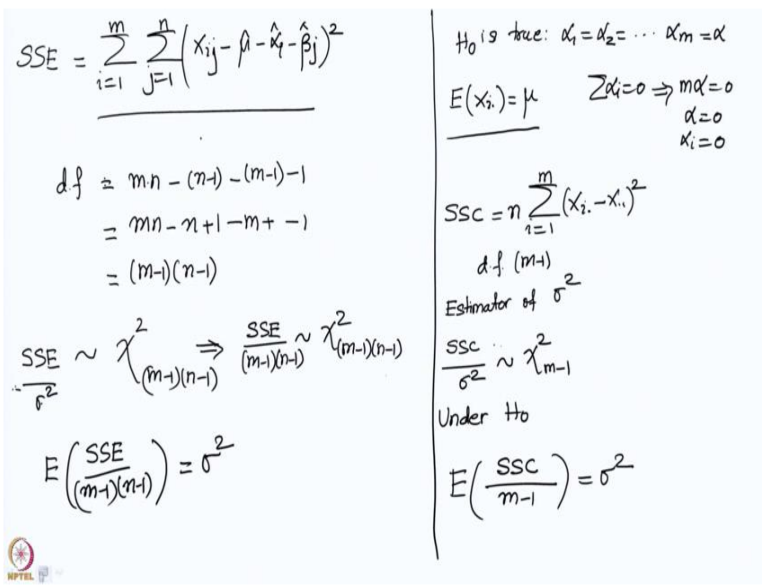
And then, this estimate is called Sums of Squares of Error. Why it is error? Well, these are your estimates and these are your actual values so you have estimated the, you have estimated the parameters through the actual values and therefore the difference between the two, the actual value and the estimated value is the error and therefore we say that this is the sum of squares of error.

(Refer Slide Time: 20:32)

$$SSE = \frac{\sum_{i=1}^{m} \sum_{j=1}^{n} \left(x_{ij} - \mu - \hat{\alpha}_i - \hat{\beta}_j\right)^2}{d.f}$$

$$\begin{aligned} d.f &= m \cdot n - (n-1) - (m-1) - 1 \\ &= mn - n + 1 - m + \ - 1 \\ &= (m-1)(n-1) \end{aligned}$$





Let us move on, so then we get the sum of squares of error, we can define as summation i is equal to 1 to m, summation j is equal to 1 to n, Xij minus Mu hat minus alpha, sorry, alpha i, alpha i minus beta j hat whole square divided by some degrees of freedom. Let us calculate this degrees of freedom.

We have totally m times n observations, Xij is m times n observation of which we have already estimated n minus 1 observations minus m minus 1 observations that is, because there are m is, so we have to subtract m minus 1, remember that in observation of alpha i you already have X dot dot and therefore it is alpha i hat is n minus 1, this is m minus 1 and then minus minus 1 for Mu hat and therefore this equals to m n minus n plusplus1 minus m plusplus1 minus 1 and therefore it is you can simplify to m minus 1 multiplied by n minus 1.

Remember, how it is calculated, alpha hat, if you, if you go back, let us go back, if you go back, alpha hat, here, alpha hat is calculated in this manner. So there is already one alpha there are n observations have to be calculated so alpha hat comes out of Xi dot, so there are n of them, and you have to take out one X dot dot, so it is n minus 1. Similarly, you have m minus 1, similarly this is becomes n minus 1 and this minus 1 is for the Mu hat and therefore from the total observation n m , m multiplied by n, you subtract that which simplifies to this.

So, this means that SSE is distributed, so then it really means that SSE is distributed as kai square with m minus 1 times n minus 1 degrees of freedom. This SSE is divided by, sorry, it should have been taken, SSE divided by degrees of freedom, this is my mistake, I correct it. You have to calculate the degrees of freedom and that is what has been calculated, so this SSE divided by its degrees of freedom is this that is SSE divided by m minus 1 multiplied by n minus 1 is distributed as kai square with m minus 1 multiplied by n minus 1 degrees of freedom.

This is your, this divided by sigma square, there is, I am sorry to make this mistake, we go back, it should have come there itself. This is sums of squares divided by, this is correct so next step is, here we make a correction. There is this divided by sigma square is a, sums of squares divided by sigma square is distributed as a kai square with m minus 1 n minus 1 degrees of freedom.

Now, when H0 is true, when H0 is true, it means that alpha 1 is equal to alpha 2 is equal to dot, dot, dot, alpha m, then expected value of Xi dot is only Mu. Why? Because summation of alpha i is equal to 0 which implies that m times if this all of them are equal to alpha, then m alpha is 0 and therefore, alpha itself is 0.

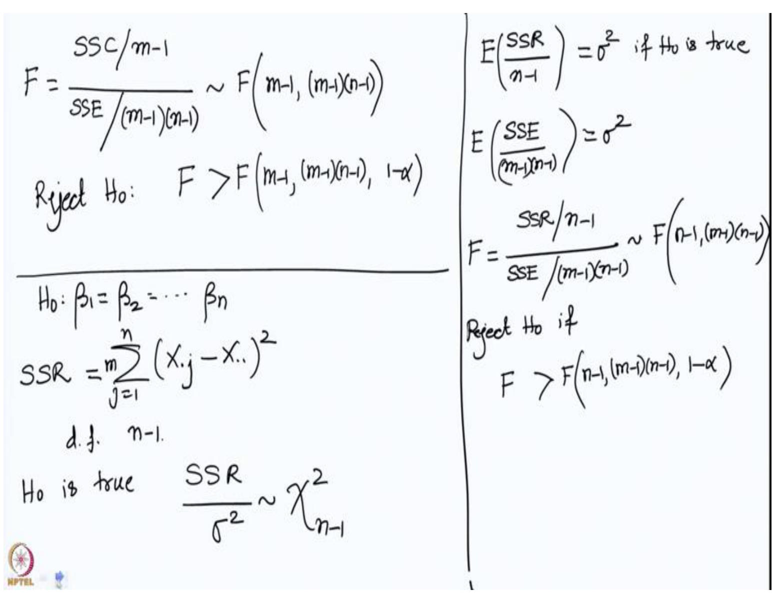
So it means that alpha ’'s are 0 and therefore Xi is equal to, expected value of Xi is Mu, so this give you sums of squares due to column, which is nothing but summation n times summation of i is equal to j, i is equal to 1 to m, Xi dot, here minus X dot dot whole square. This is now within the column, so we found that the total sums of squares sits here where you have taken the full value, where we have taken the full, the complete, all the values Xij and their difference from the mean value that is called the sums of squares of error.

Now you do it only for the columns, so it is called sums of squares by the column, so it is sums of squares by column, you remember we had a within group sums of squares and between group sums of squares, this is a sums sums of squares between the columns and therefore it is Xij minus X dot dot whole square multiplied by n.

And therefore, this has a degrees of freedom of m minus 1 and this is also an estimator of sigma square. Why? Because SSC divided by sigma square is distributed as kai square with m minus 1 degrees of freedom. Here also, I stand corrected myself. I think there are many errors I have made. This was correct, here instead of degrees of freedom, it should have been sigma square. So then this is distributed as kai square m minus 1, so you have now expected value of SSE divided by m minus 1 multiplied by n minus 1.

This is your one estimate of sigma square and under H0, expected value of SSC that is sums of squares by column, column sums of square, so between the column what you take as a sums of squares divided by its degrees of freedom m minus 1 is also sigma square and therefore you can say that the ratio of this would give you an a test statistic to see if the ratio, the thing is correct.

(Refer Slide Time: 29:48)



So then we go to the next step, so we have, if we take that, the ratio, I call it again F, as sums of squares due to column, divided by its degrees of freedom, divided by sums of squares of error, divided by its degree of freedom, is distributed as F with m minus 1, m minus 1 times n minus 1 degrees of freedom and you reject the null hypothesis, you reject H0. If this F is greater than F m minus 1, m minus 1 times n minus 1, 1 minus alpha is true, then you are going to reject the null hypothesis, otherwise you are going to accept the null hypothesis. So this is how it is done.

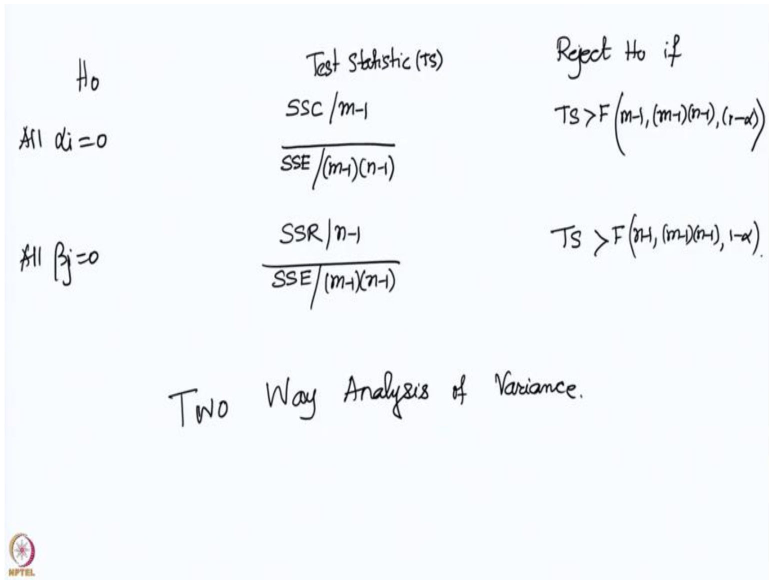
Suppose you take the another null hypothesis that beta 1 is equal to beta 2 is equal to etc , beta n, then naturally you are going to calculate what is known as SSR, row between the row sums of square. So again, against the row you are going to make the sums of squares, so that is going to be summation, that will look like a summation of m times j is equal to 1 to n Xj, X dot j minus X dot dot whole square, its degrees of freedom will be n minus 1 and you have S, this will be also when, when H0 is true, you will find that SSR divided by sigma square is distributed as kai square with n minus 1 degrees of freedom.

So again, you we have two estimate of sigma square, SSR divided by its degrees of freedom, sorry, it should be n minus 1, n minus 1 is equal to sigma square if H0 is true. Of course, we have expected value of error sums of squares divided by m minus1 times n minus 1 is also sigma square.

So the F ratio which says that sums of squares due to rows, divided by its degree of freedom, divided by error sums of squares divided by its degrees of freedom, is distributed as F with n minus 1, m minus 1 times n minus 1 degrees of freedom and therefore, we have, we say that reject null hypothesis if this F is greater than F n minus 1, m minus 1 times n minus 1, 1 minus alpha, this is how it is done. So the two hypothesis can be tested.

Let us, I would like to write the whole thing down in one table so that it becomes easier for us to understand.

(Refer Slide Time: 34:04)



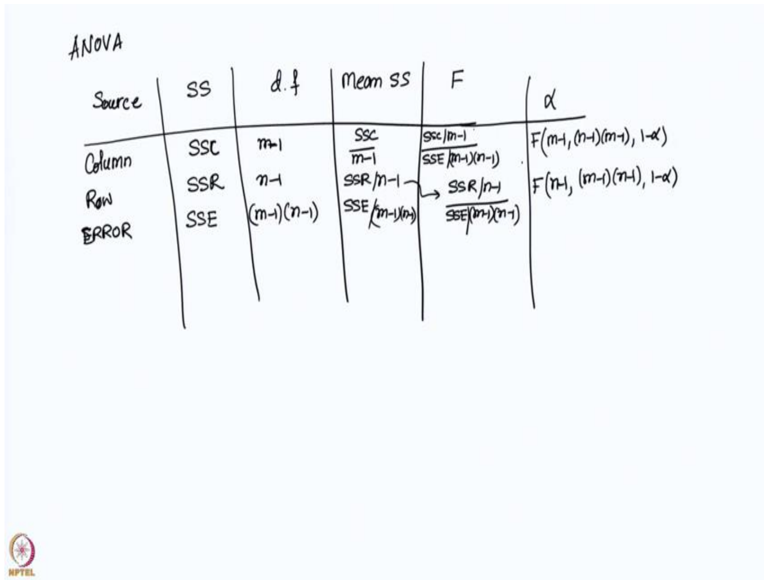
Let us consider different hypothesis, the test statistic and the rejected, reject H0 if, so if you take H0 as all alpha i is equal to 0, the test statistic is sums of squares due to column, divided by m minus 1, divided by sums of squares of error, divided by m minus 1 times n minus 1, you call this F, this is a T statistic, I call it a TS and you reject if TS is greater than F m minus 1, m minus 1 times n minus 1 degrees of freedom and 1 minus alpha.

If you want to test that all beta i is equal to beta j equal to 0, then you are looking at SS row divided by n minus 1, divided by sums of squares of error, sums of squares of error divided by m minus 1 multiplied by n minus 1 and then your test statistic has to be greater than F n minus 1, m minus 1 times n minus 1, 1 minus alpha.

This is how it is to be calculated, this is called Two Way, sorry, Two Way Analysis of Variance. So let us summarize it. It is important that we simplify this problem into two components of the columns and rows and then we test the hypothesis that all columns are equal, having the equal mean or all rows having an equal mean.

Both of them convert themselves into having a estimate of way population variance under the null hypothesis, when the null hypothesis is true, it gives you a variance of, estimate of a variance.

(Refer Slide Time: 37:21)



While if you take the total sums of squares, so if you look at the table, there is something called a table and ANOVA table is says that sums, the source, due to which you are considering the sums of squares, the sums of squares themselves, the degrees of freedom, the mean sums of squares, and the F ratio. So the source would be like SSC, that is, it will, sorry, the source will be called the column so it will be column. The sums of squares will be SSC, degrees of freedom will be n minus, sorry, m minus 1, then it will be SSC over m minus 1 this is your thing.

Then there will be row rows. So you have sums of squares due to rows, it will have n minus 1 degrees of freedom and then, SSR over n minus 1 is mean sums of squares and then you will have sums of squares due to error which will be called SSE, which has a degrees of freedom n minus 1 times n minus 1. And you have SSE divided by m minus 1 times n minus 1 and for here the F statistic would be SSC by m minus 1, divided by SSE by m minus 1 times n minus 1. And in this case it will be SSR by n minus 1 divided by SSE m minus 1 divided by n minus 1.

The large value of F, you can work out the probability of F or you have an alpha and then the large value of F actually gives you the respective cut off value so here you can have cut off value as m minus 1, n minus 1 times m minus 1, 1 minus alpha. Here it will be F n minus 1, m minus 1 times n minus 1, 1 minus alpha. So you know that if the value, this F value is larger than this value, then you are going to reject the null hypothesis otherwise you are going to accept it. This is called Two Way Analysis of Variance and this is called Analysis of Variance table. Thank you.