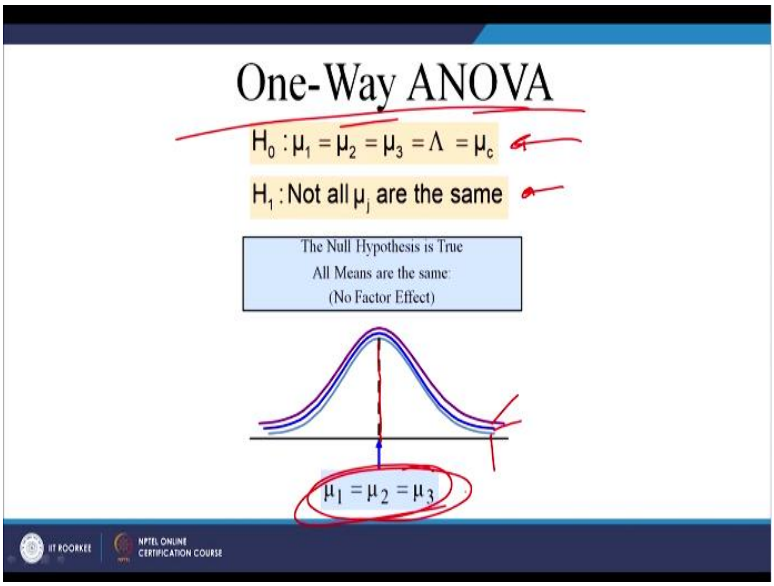
Business Statistics
Prof. M. K. Barua
Department of Management Studies
Indian Institute of Technology – Roorkee

Lecture – 46
Analysis of Variance - I

Hello friends. I welcome you all in this session. As you are aware, in previous session we discussed about design of experiment and analysis of variance. In this session, we will continue our discussion on analysis of variance. So let us look at one-way ANOVA.

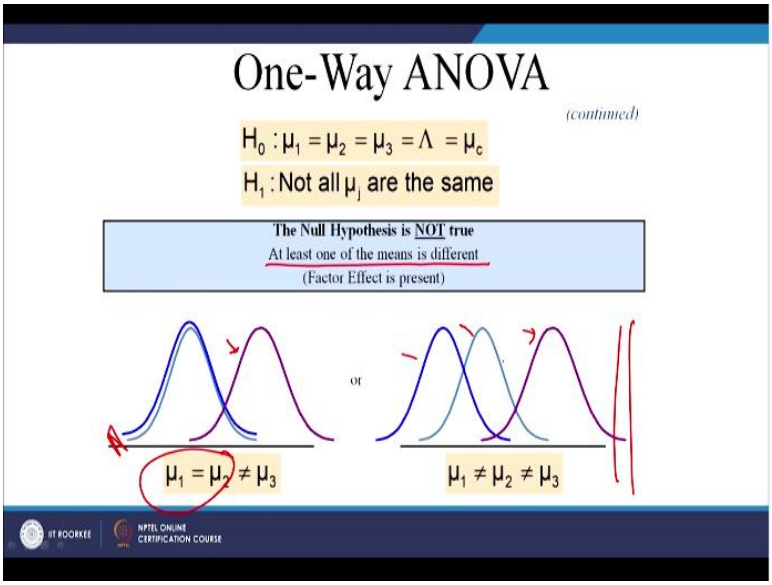
(Refer Slide Time: 00:46)



One-way analysis of variance, as I have already told you that one-way analysis of variance is used for analyzing completely randomized design. So null hypothesis is that all populations means are same and they are not same is alternative hypothesis. So let us look at these 3 distributions, 1, 2 and 3. So these are 3 distributions having same mean right. So we want to know are these means same or is there any significant difference amongst these means.

So that is to be tested. We are not bothering anything about variances or standard deviation. We are bothering about means only.

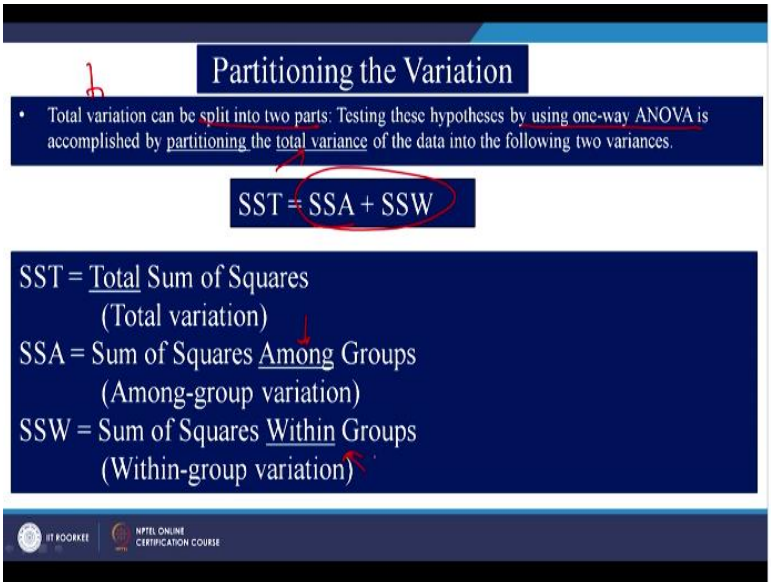
(Refer Slide Time: 01:41)



So this is how you can represent these 3 distributions. So null hypothesis, the null hypothesis is not true if at least one of the means is different. So let us say if μ1 and μ2 are equal. Let us write μ1 and μ2 are equal but μ2 and μ3 are not same, will reject null hypothesis or you may have let us say μ1, μ2, μ3, all 3 are different from each other okay. So this is how you can represent distribution.

So these two means are same but this mean is different but all these 3 means here are different from each other.

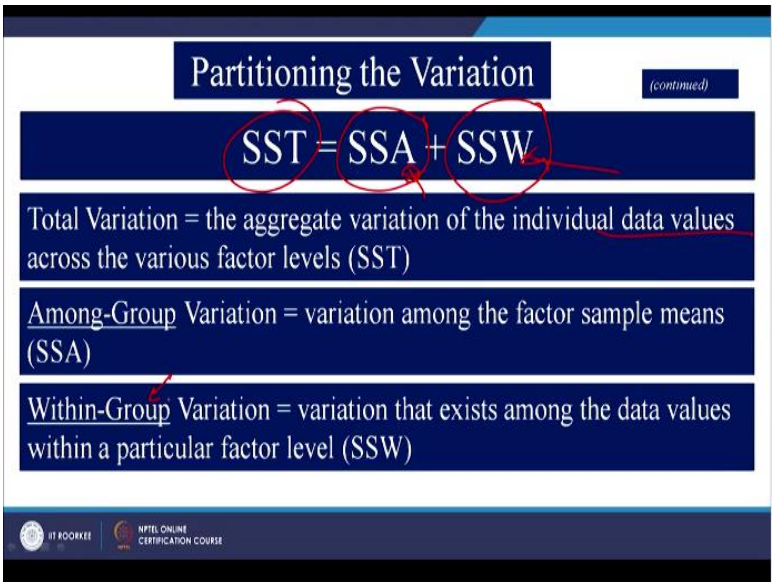
(Refer Slide Time: 02:27)



So as I said there is something called total sum of squares which is nothing but the summation of this, these two, the Among-group sum of squares and within-group sum of square. So total variation it can be partitioned or can be split into two parts using one-way

ANOVA is accomplished by partitioning the total variance and this total variance is nothing but among-group variance and within-group variance okay.

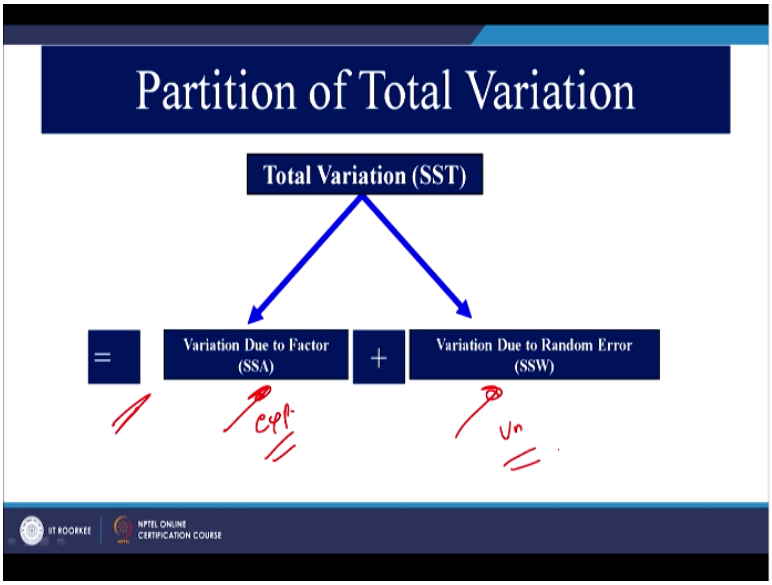
(Refer Slide Time: 03:08)



So this is total sum of square which is sum of among-group sum of square and within-group sum of square right, 'W' stands for within and 'A' stands for among right. So total variation is the aggregate variation of the individual data values across the various factor levels. Among-group variation, variation among different groups right. So let us say there are 3 groups, so difference among first second, second third and first and third.

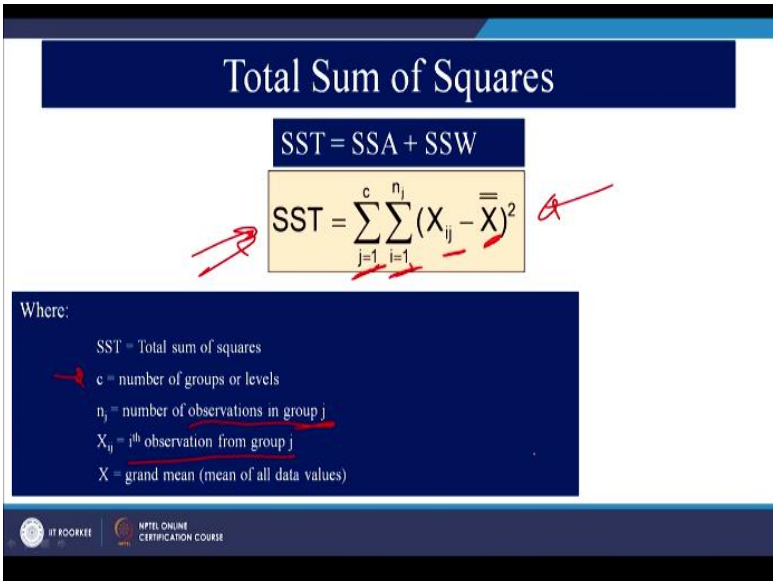
Within-group variance, you will always have some variance within column or within-group because you will always have certain elements in each group and there would be some variance within those elements. So that is called within-group variation.

(Refer Slide Time: 04:02)



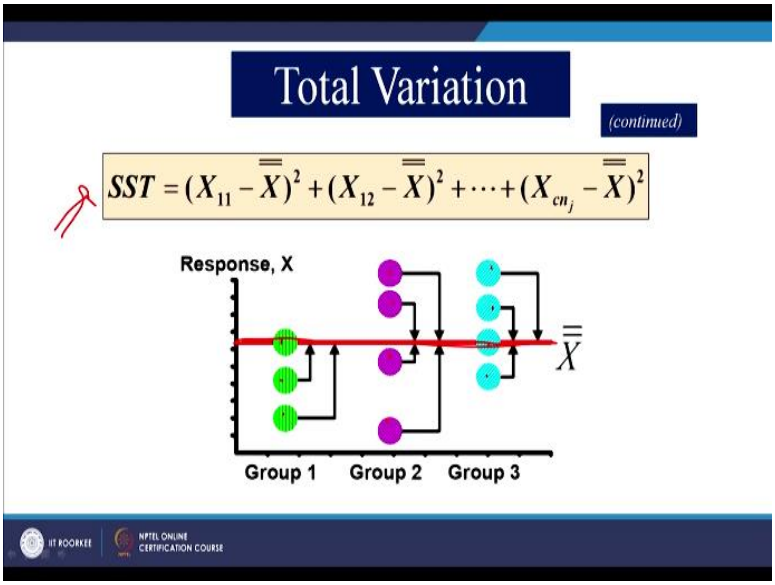
So total variation is sum of variation due to factor and variation due to random error right. So this is unexplained variance and this is explained variance, is not it? This is what we have discussed earlier as well.

(Refer Slide Time: 04:21)



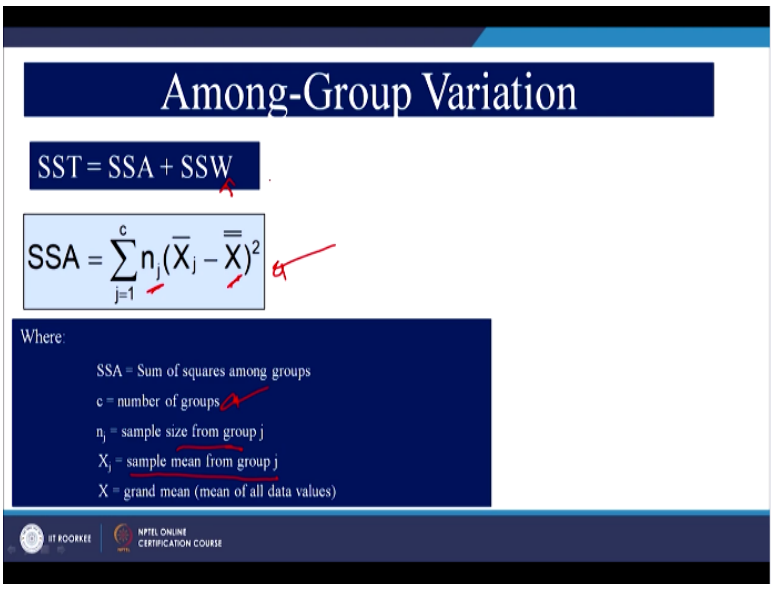
So total sum of squares can be calculated like this. It is the summation of all j ranging from 1 to c, all i's ranging from 1 to nj. Here nj is number of observations in each group j. Xij is nothing but ith observation for jth group, X double bar is grand mean right. So c is number of groups or number of levels. So this is how you should be calculating total sum of squares.

(Refer Slide Time: 04:59)



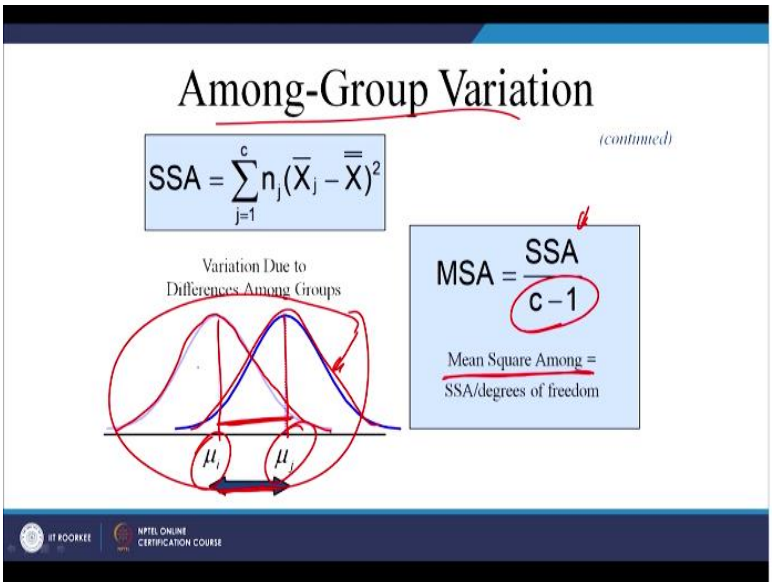
So total variation let us say if there are 3 groups and these are different observations in these 3 groups, so first, second, third; first, second, third, fourth; first, second, third, fourth right. So 3 observations in first group, 4 observations each in group 2 and group 3. So this is how you can calculate total sum of squares. So first of all you would be calculating group mean of all these 3 groups and then this is nothing but grand mean right X double bar which is this line okay.

(Refer Slide Time: 05:42)



Let us look at among-group variation; this is how you can calculate Among-group variation. So this is n_j is number of sample size from group j, X_j is sample mean from group j, X double bar is grand mean and c is number of groups right. So if we know Among-group variation and if we know within-group variation, we can add these two variations to get total sum of squares. So this is Among-group variation.

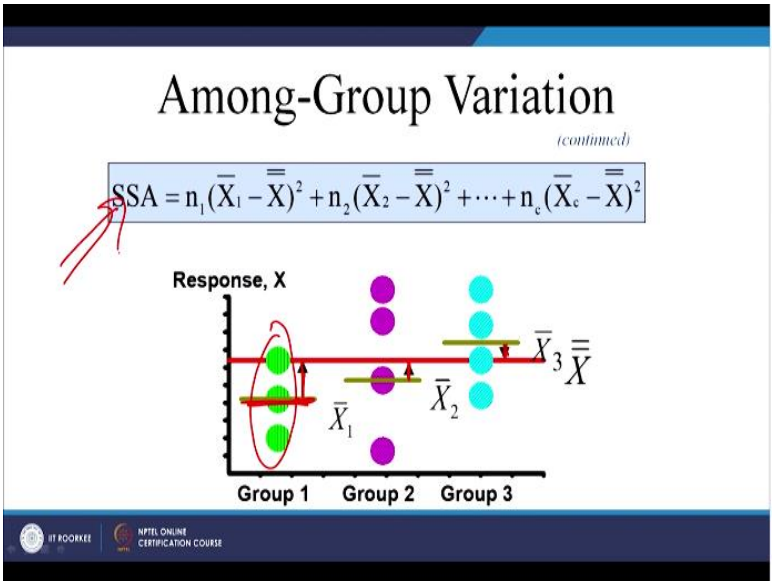
(Refer Slide Time: 06:25)



Now this is how among-group variation looks like right, so there is one one distribution right. Its mean is this okay. Another let us say there is one more group whose distribution is this. So this is the mean of second group. So this is among-group variation right, this is what is the variation between these two groups okay. So mean sum of square, mean square among is sum of square among group/c-1, here c is what number of groups right okay.

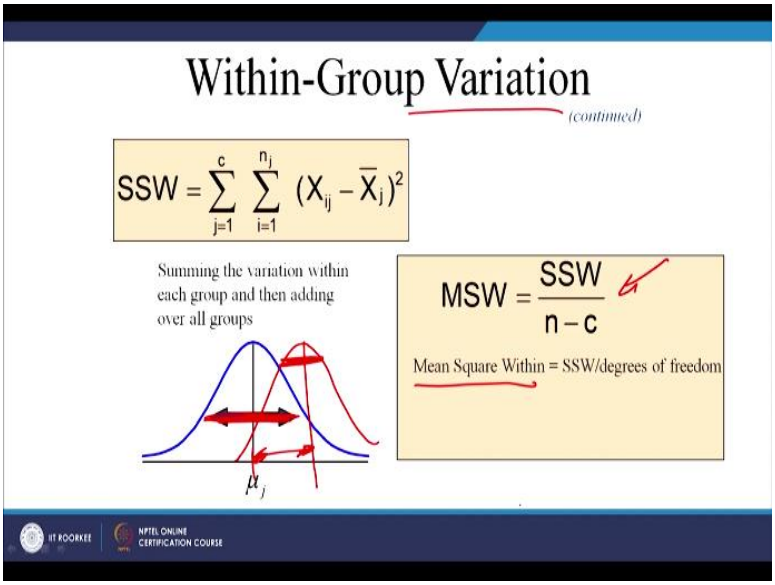
So this is how among-group variation looks like or in this case since there are two groups, will call it between group variation right. If you have got more than two groups, will call it among-group variations. So this is an among-group variation.

(Refer Slide Time: 07:36)



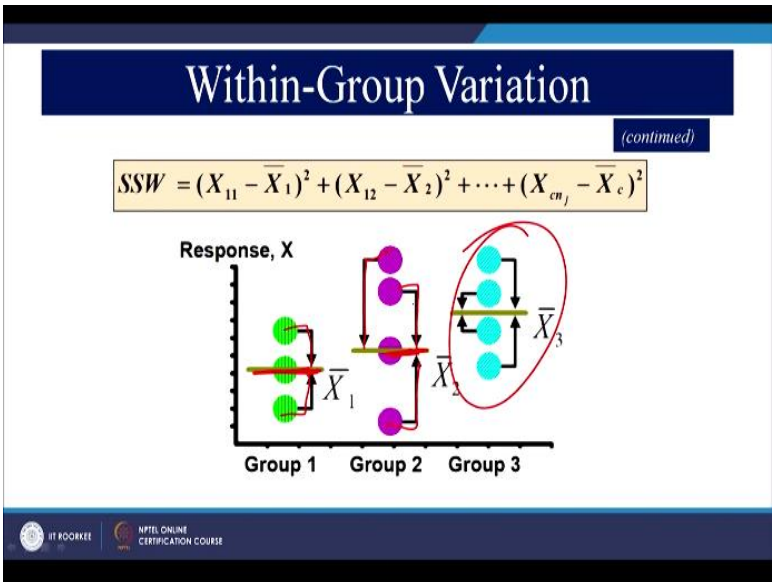
So among-group variation would look like this. So you have got 3 elements in first group, so this is the mean of first group and how that mean is away from grand mean right. So this is the difference, this is the difference and this is the difference right. So this is among-group variation okay.

(Refer Slide Time: 07:59)



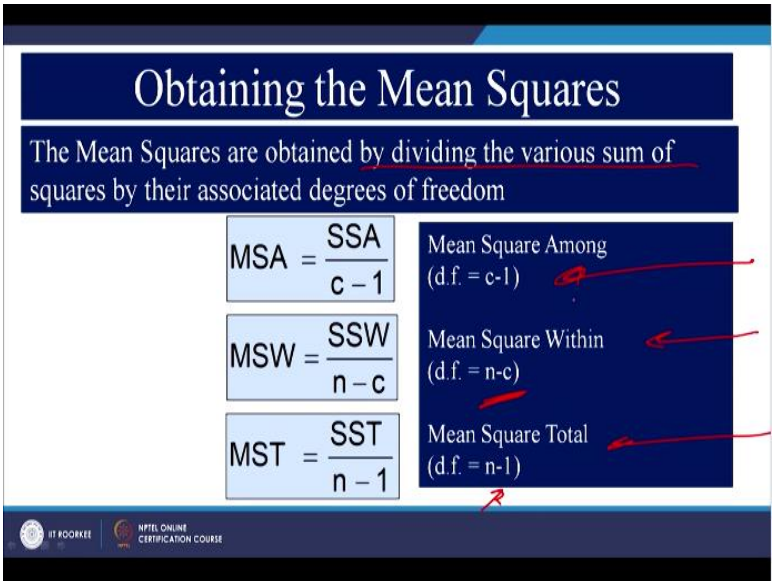
Let us look at within-group variation. So for within-group variation, this is the formula you can use and all these terms are there. We have already defined them. So within-group variation and this is how within-group variation looks like right. So this is the variation within-group and Among-group how it was, this is among or between group and this is within-group. For example, this is within-group variation right okay. So the mean square within can be calculated by this formula right.

(Refer Slide Time: 08:48)



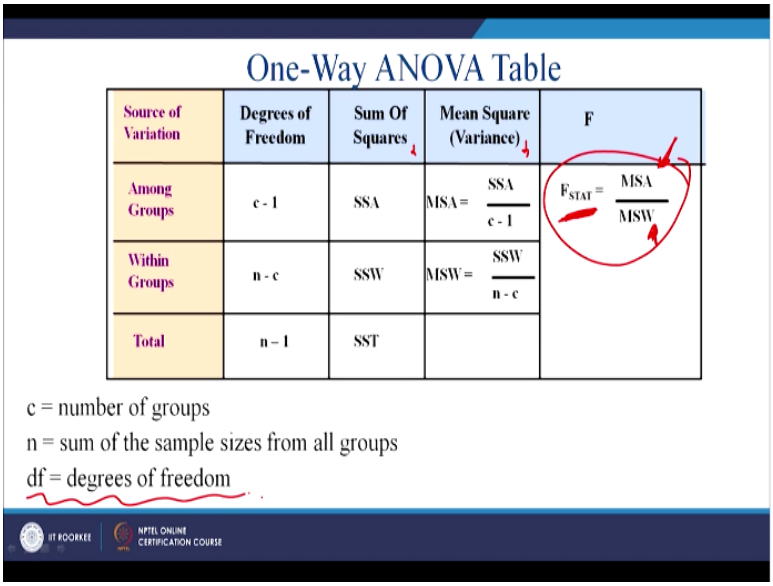
Within-group variation, so these are different elements, so variation from mean of each element. The variation of each element from mean of the group, so this is the mean and this is the difference. This is the mean of second group and these are the variation from mean of each item or each group member, similarly for third group.

(Refer Slide Time: 09:23)



So this is how you can calculate within-group and among-group variations right. So the mean squares are obtained by dividing the various sum of squares by their associated degrees of freedom. So the degree of freedom for total is n-1, for within-group it is n-c and for among-group it is c-1 okay. What is c? C is number of groups right.

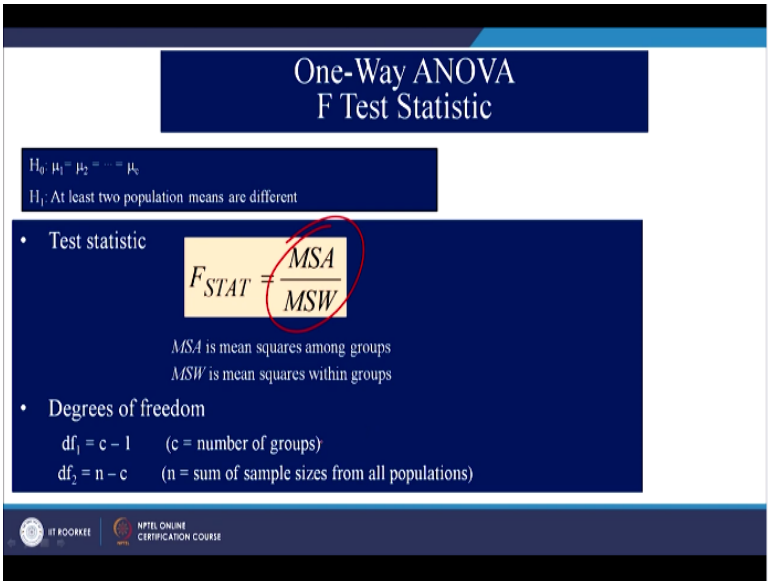
(Refer Slide Time: 09:59)



So this is how the one-way ANOVA table looks like. The output from let us say excel or Minitab or any other software. So you will have degrees of freedom, sum of squares and

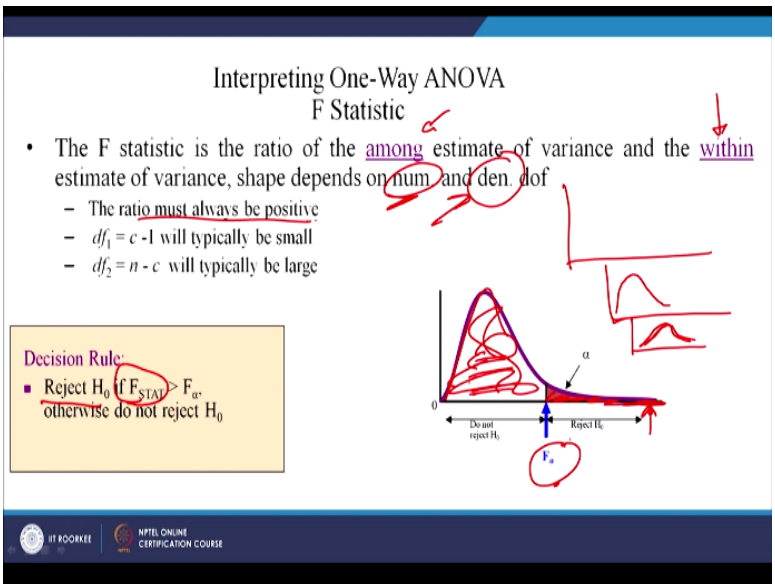
mean square and then F statistics. So F statistics is nothing but it is the ratio of among-group variation and within-group variation. So we will always compare among-group and within-group variation and we will see whether this F statistics is within rejection region or acceptance region okay. So we have already talked about this df is degree of freedom right.

(Refer Slide Time: 10:53)



So F statistics is ratio of among-group to within-group with these degrees of freedom.

(Refer Slide Time: 11:08)

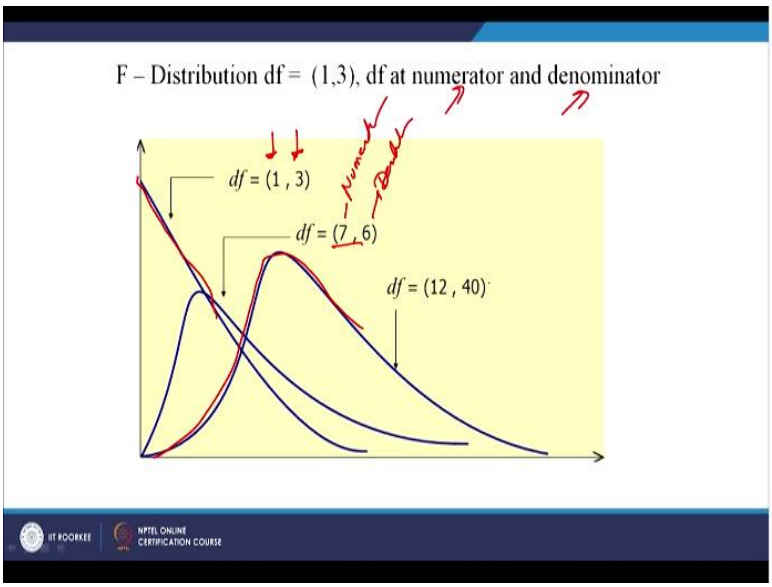


So if you look at F statistics or F distribution, F distribution is always one-tailed test, keep in mind. It is not a two-tailed tests; it is always one-tailed test. So this is your rejection region in F statistics and this is non-rejection region right this side this entire region is non-rejection region. So the F statistics is the ratio of among-group and within, among estimate of variance and within estimate of variance.

And the shape of this distribution depends on the number of degrees of freedom of numerator and denominator. So let us say this is shape of distribution, if you increase numerator degrees of freedom and denominator degrees of freedom, then this shape would approach normal distribution right. It will become a normal distribution.

So the ratio must always be positive and the decision rule is this if this is your critical value, if critical value is let us say less than or if critical value is less than calculated value and if it is in this region calculated value, then will reject the null hypothesis. If it is in this region, will not reject null hypothesis right. So will not reject null hypothesis if F statistics is greater than alpha right.

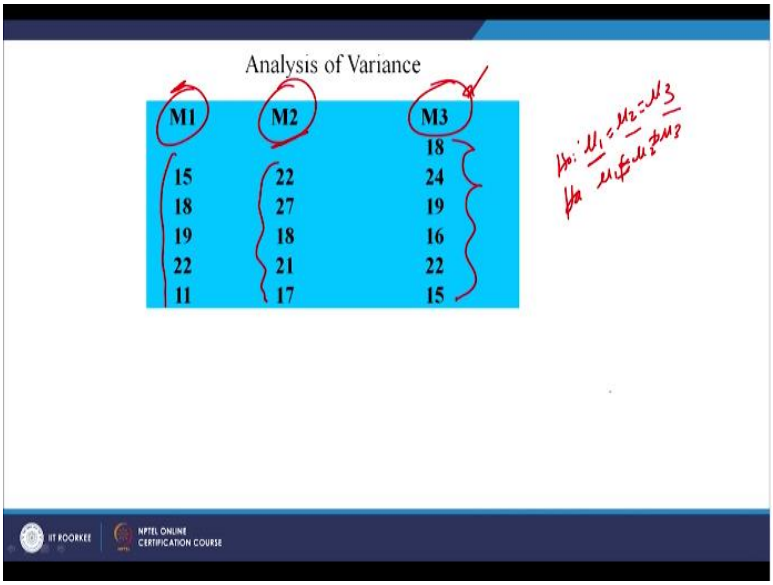
(Refer Slide Time: 12:54)



So this is how the shape of F distribution changes with change in degree of numerator of, with degree of freedom of numerator as well as denominator right. So this is the shape of the curve right for 1 degree of freedom numerator and 3 degree of freedom denominator right. So the first one would always be for numerator and the second one is always for denominator right.

So just see how this shape is changing right with change in degree of freedom of numerator as well as denominator.

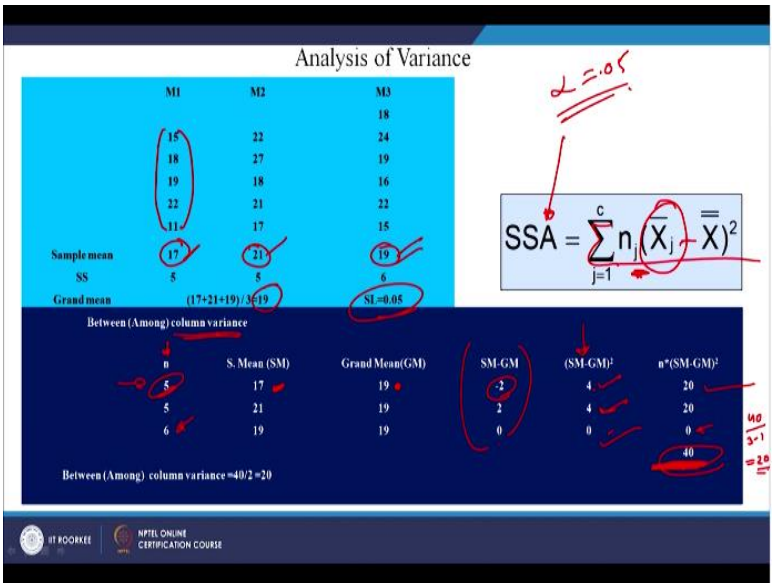
(Refer Slide Time: 13:33)



So let us look at this question on analysis of variance. So let us say there are 3 different teaching methods, M1, M2 and M3 and these are the marks of 5 students when they were taught by using method 1, marks of 5 students when they were taught by method 2 and these are the marks of 6 students when they were taught by method 3 and the question is, is there significant difference in effectiveness of these 3 teaching methods?

So how will you frame null and alternative hypothesis? So first of all, you will say that μ1= μ2= μ3 right. Mean of method 1, method 2, method 3 all these are same, so this is your null hypothesis right. Alternative hypothesis, they are not same right. So let us say mu 1, mu 2, mu 3 they are not same or at least one of them is not same okay. So first of all you need to find out among-group variations and within-group variations right.

(Refer Slide Time: 15:11)



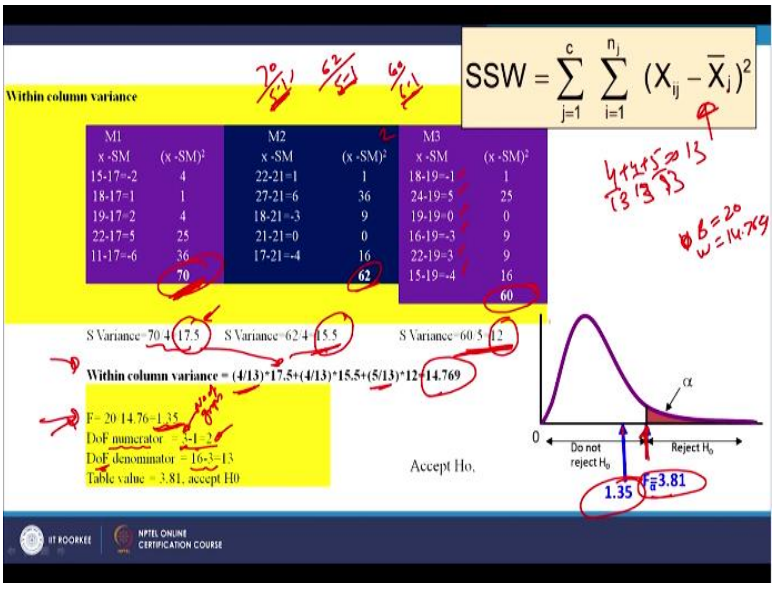
Let us look at this question. So first of all, you will find out among-group variation right, not sum of squares among-group because there are 3 groups, will call it among-group. So among-group variance are column variance right. So first of all what you should do? Find out the sample mean of each of these groups. So summation of all these is just add all these 5 numbers divided by 5 will get 17, here 21 and here 19.

After finding out mean of each of these groups, find out grand mean which would be mean of these 3 means. So 17, 21, 19/3, so 19 is the grand mean and we have to test this hypothesis at 0.05 significance level right. Now let us calculate among-group variance right. So first sample mean of first group, sample mean is this, sample mean is grand mean, sample mean-grand mean is -2, 17-19-2, 21-19+2, 19-19 0, so sample mean-grand mean take the square of this column which is 4, 4, 0.

And multiply this by the sample size or the number of units in each of those samples right. There are 5 units in first simple, so 4*5=20, 4*5 again 20, then 6*0 6 and the total, this is what we have written here, summation of nj right, nj is the number of items in that particular group right*mean-grand mean whole square right, so this is 40. So SSA is 40, so between column variance would be just divided by c-1, here c is nothing but number of columns right.

So there are 3 columns, 3-1 is 2, so that is why we are dividing 40/3-1 which is=20. So we have calculated between column variance.

(Refer Slide Time: 17:59)



Now let us look at within column variance, for within column variance this is the formula you have to apply but let me tell you how to find out within column variance. Now when I say within column means in column itself right, between column means between these two columns right M1 and M2 between M2 and M3 right. So this is this between column and this within column right.

So we will now find out within column variance. Since we know the mean of first group which is 17, so the first element in first group is 15, second element is 18, third 19, fourth 22, fifth 11. So subtract sample mean from each of these elements and take a square of it. So this is 4, 1, 4, 25 and 36, add all these right. So this is what $(X_{ij} - \overline{X_j})^2$ which is this, summation of this is 70 right.

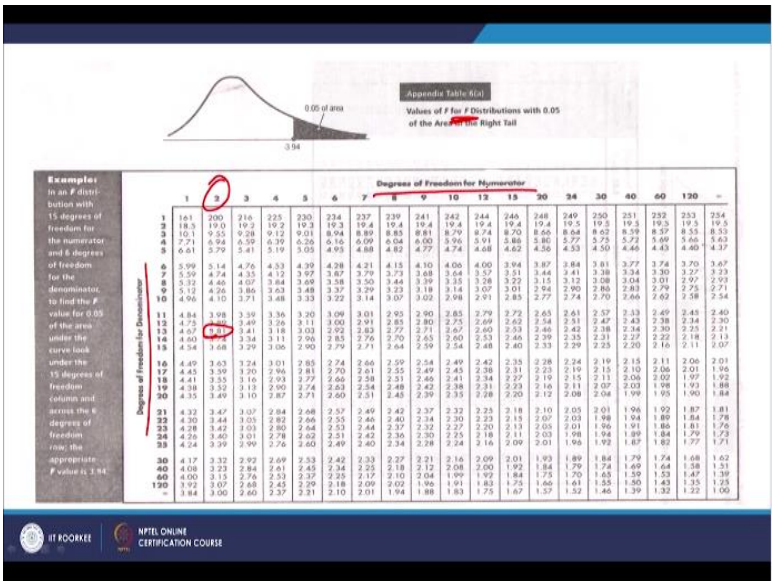
Similarly, for second column the sample mean of second column is to be subtracted from all the elements of second group or second column. Take square of it and summation of it, so 62. Similarly, for M3 or third group is 60 right. So now the variance is 70/4. Why 4? Because there are 5 items right so 70/5-1 this is 62/there are 5 members or 5 units in group 2 right 5-1. Then 60/how many samples here, 1, 2, 3, 4, 5, 6 right 6-1 right.

So 70/4, 62/4 and 60/5 so this is 17.5, 15.5 and 12. So within column variance would be like this, it is 4/13*17.5. Why this 4/13? 13 is 4+4+5 is 13 right, this 4, this 4 and this 5 so total is 13. So 4/13, 4/13 and 5/13 so 4/13*this and 5/13*12. So this is how you would be getting as within column variance. So between column variance was 20 and within column variance is 14.769.

So the F statistics or F value would be 1.35. So this is the calculated value of F. Now this is here right, now we want to find out F critical. So this is to be we will have to check F table for F critical value and we have to check F value at 3-1 degree of freedom of numerator. Why 3-1? Because there are 3 groups right, so number of groups-1 is=2 and for denominator what we want, it is 16-3.

How come it is 16? There were total 16 members in these 3 groups; 5, 5 and 6 right, will check it again right, 5, 5 and 6 so total 16, 16-number of groups and this is number of groups-1 okay, so 2 degrees of numerator and 13 degrees of freedom of denominator right.

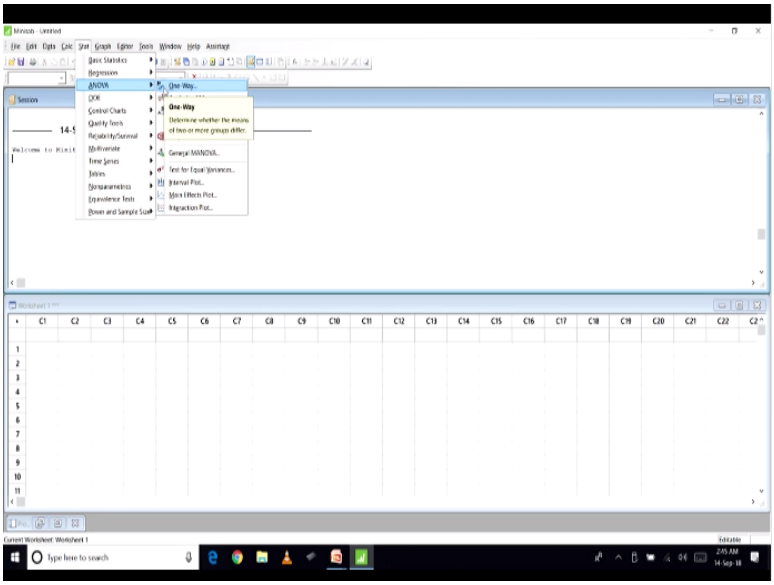
(Refer Slide Time: 22:52)



So let us look at table, 2 numerators so we have got this F table right. So you have got degrees of freedom for numerator, degrees of freedom for denominator. So at two degrees of numerator, two degrees of freedom for numerator and 13 right, 13 degrees of freedom for denominator. So this value is 3.81 which is here 3.81. So the calculated F value is in non-rejection region, will not reject null hypothesis.

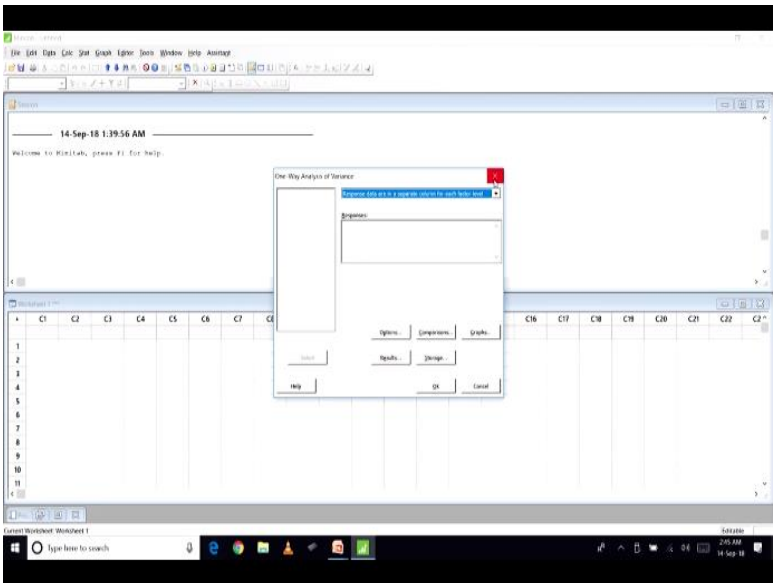
So our conclusion is will not reject null hypothesis and what was it our null hypothesis, let all these 3 means are same. So will say that all these 3 methods are equally effective, there is no significant difference amongst these methods. Now let us look at the same question using Minitab, so will solve this question using Minitab and will see the critical F value, will look at P value and so on right okay.

(Refer Slide Time: 24:10)



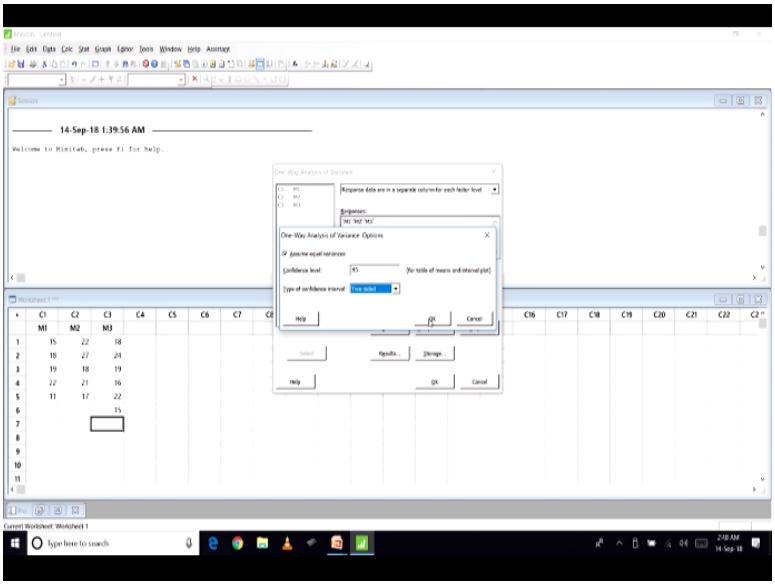
So we will go to Minitab software, so will go to stat, will go to ANOVA will go to one-way ANOVA.

(Refer Slide Time: 24:19)



So response data are in column for all factors or response data are in separate column for each factor okay so will give data entry.

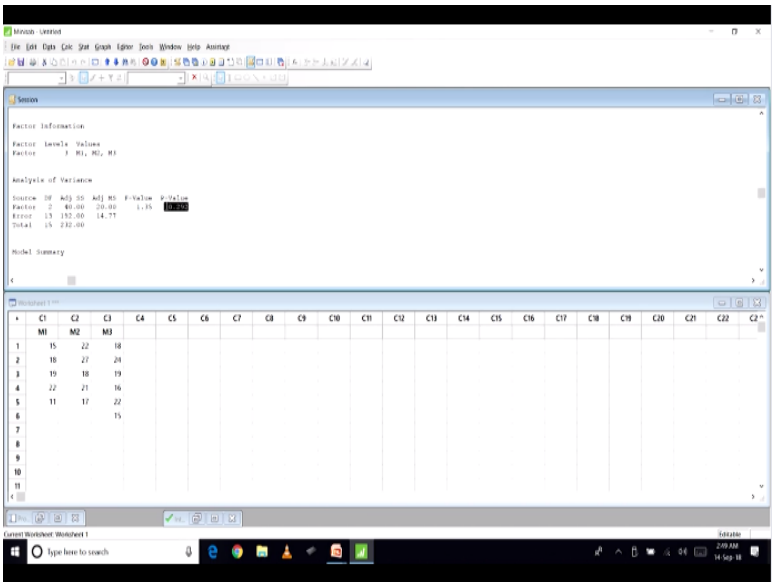
(Refer Slide Time: 24:36)



So this is M1, M2 and M3, 15, 18, 19, 21 I think, 22, 11 right. So this is the response values or dependent variable values for first method. Then, 22, 27 and 18 then you have got 21 and 17 right but for the third one we have got 6 elements right so 18, 24, 19, and then we have got 22, we have got 16, 22 and 15 right. Let us look at we have to test this hypothesis 0.05 significance level right.

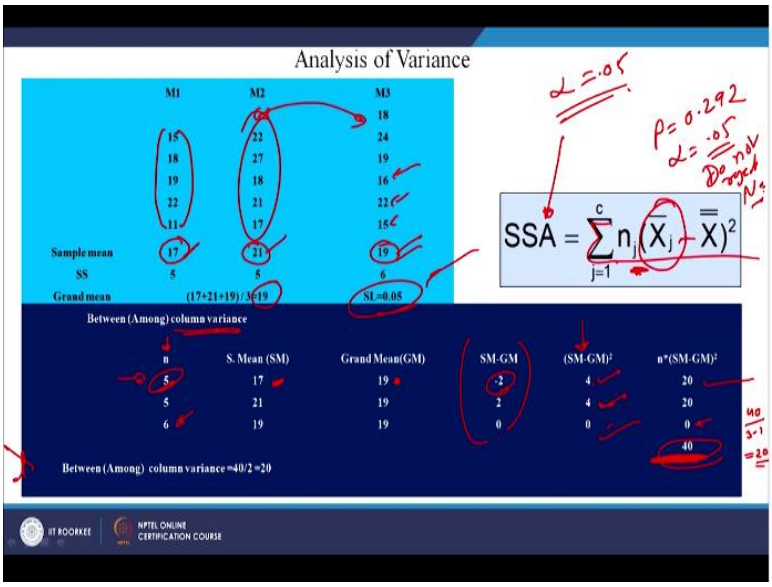
So let us look at how to enter the significance value. So one-way ANOVA, so response data are in separate column for each factor level okay, so we will select these 3, so select responses 1, 2 and 3, will go to options. So confidence level is 95%, the type of confidence interval is just click over here, so it has to be in upper tailed test or in fact so let us look two-sided okay.

(Refer Slide Time: 26:52)



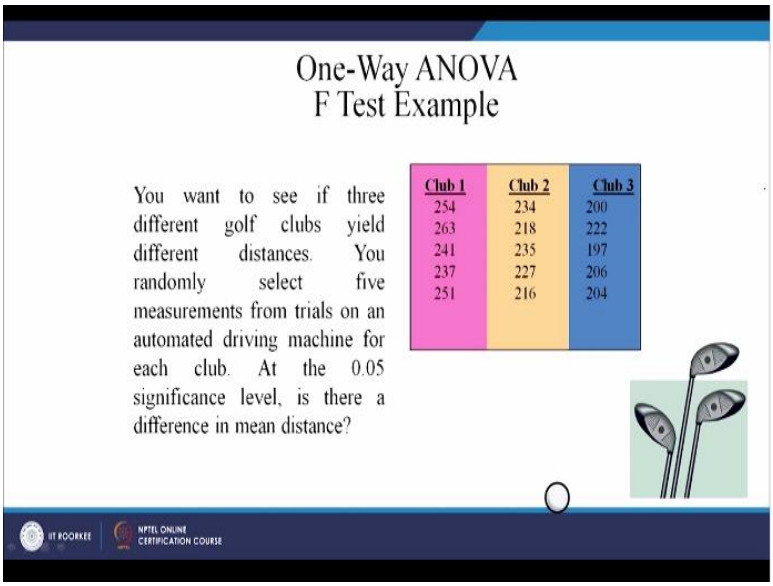
Let us look at results, what results we are looking for. Display okay simple tables we want so we want all these things and we want means, we want ANOVA, model summary and everything we are looking for right, so will click okay. So if you look at this table, in fact will not analyze much this table, will just look at P value. So if you look at P value here is this is 0.292 and alpha is=0.05.

(Refer Slide Time: 27:46)



So P value is not less than alpha. So we do not reject null hypothesis. So you are getting the same answer either you do it manually or you do it through Minitab software. So this is how you can solve this question.

(Refer Slide Time: 28:20)



In next class, we will work out this example using Minitab software and will try to read some more information from Minitab output table. So for the time being let me stop here. Thank you very much.