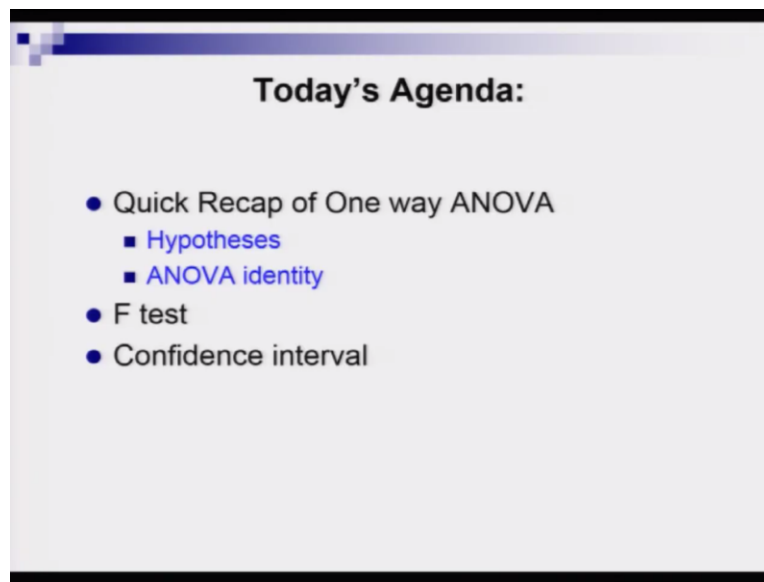


**Applied Statistics and Econometrics**  
**Professor Deep Mukherjee**  
**Department of Economic Sciences**  
**Indian Institute of Technology, Kanpur**  
**Lecture-18**  
**One-way ANOVA**

Hello, friends. Welcome back to the lecture series on Applied Statistics and Econometrics. So, today, we are going to continue our discussion on one-way ANOVA. So, as I told you previously that ANOVA is a very interesting field where you will find bits and pieces of hypothesis testing and linear regression analysis. So, today also we are going to stay in the field of or in the area of one-way ANOVA only, but we are going to complete the discussion on it. So, let us have today's agenda item.

(Refer Slide Time: 0:49)



So, we are going to start with a brief recap of one-way ANOVA. I think it is not a bad idea to go through some basic concepts for better continuity, because ANOVA is a bit complicated concept. Here, in this lecture, I am going to remind you about the hypothesis that we have to frame, then I am also going to remind you about the ANOVA identity, and we are going to talk about F test at length. And finally, we are going to conclude today's lecture by having a very brief discussion on confidence interval.

So, what is ANOVA? If you remember, we were discussing, partitioning or decomposition of total variation in  $y$ , the dependent variable, in two parts. The one part is systematic factor and the other party is the idiosyncratic error factor or random factor. Now, what is the systematic factor? Systematic factor actually tells you what is the variation in the data from the existence of different groups. So, if you put different  $y$  values in different buckets for different groups, then there will be sample mean related to those groups.

So, there will be a dispersion between these sample means, so how that is going to affect the overall variability in the data. Is there any way we can establish that this grouping variable has an impact on the original dependent variable  $y$ . So, that is basically the question with which we have started our journey. And systematic error is basically taking care of that part of the story, that whether a grouping variable, which is a qualitative variable with more than two levels, whether that has any impact on the variation of  $y$ .

(Refer Slide Time: 2:57)

### The Model

- Hypothesis formation:
  - $H_0 : \mu_1 = \mu_2 = \dots = \mu_k$
  - $H_1 : \text{Not all population means are the same}$
- The model for the observed response is given by:  $Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$ 
  - $Y_{ij}$  = Variable value for item number  $j$  in group number  $i$
  - $\alpha_i$  = effect of being in group number  $i$
  - $\varepsilon_{ij}$  = idiosyncratic error  $\sim N(0, \sigma^2)$
- Need to estimate all  $\mu$  and error variance  $\sigma^2$
- Total variation in  $Y$  or Total Sum of Squares (SST) can be split in:
  - ✦ Sum of Squares Between/Among groups (SSB)
    - = dispersion between the factor/group sample means
    - ↔ Between (k groups)
  - ✦ Sum of Squares Within groups (SSW)
    - = dispersion that exists among the data values within a particular factor level
    - ↔ Within (individuals per group)

So, in this slide, I am going to remind you quickly about the null and alternative hypothesis. So, we start with  $k$  number of population groups. Of course,  $k$  is an arbitrary number, but generally for ANOVA let me tell you that we deal with more than or equal to three groups. Now, we start with the null hypothesis that all the population means are equal and we test that null hypothesis against the alternative hypothesis, which says that not all population means are identical. So, that

means that at least one of these population means is different from the rest of the population means.

So, the next bullet point shows you the mathematical model, the equation. And that says that any particular observation  $y_{ij}$ , so here let me remind you again,  $i$  is basically for the group, it is denoting group, so  $i$  will take values from 1 to  $k$ . And  $j$  is basically the number of observations in a particular group. So,  $j$  will take values from 1 to  $n_i$ , where  $n_i$  is basically the maximum number of observations in the sample for group  $i$ .

So, a particular observation  $y_{ij}$  can be broken in two parts and it's an additive relationship. So, the first part of the sum will be the overall mean that is denoted by  $\mu$ . So, that is coming from the entire data set of  $y$ . And then  $\alpha_i$  is basically the effect of being in a group numbered  $i$ . So, that basically talks about a fixed kind of effect or systematic effect that is present in all observations. And then  $\epsilon_{ij}$  is basically the idiosyncratic error which follows a normal distribution with constant variance. So, that is basically the setup.

Now, after this expression of model equation, we have also spoken about the ANOVA identity. And ANOVA identity says that the total variation in the dependent variable  $y$  can be partitioned or decomposed in two components, namely the sum of squares between groups and sum of squares within groups. Now, the between groups and within groups, they may be a bit confusing to hear, but let me explain what are they in simple language.

So, here in this slide, concentrate on the last bullet point. Here, I am showing you the definition for SSB and SSW. So, the sum of squares between or among groups is defined as the dispersion between the factor of group sample means. So, what do I mean by this? So, suppose you start with  $k$  number of groups, and then for  $k$  number of groups you can draw  $k$  independent samples. So, from these  $k$  independent samples you can now compute the corresponding sample mean, so you will end up getting  $k$  number of sample means for  $y$ .

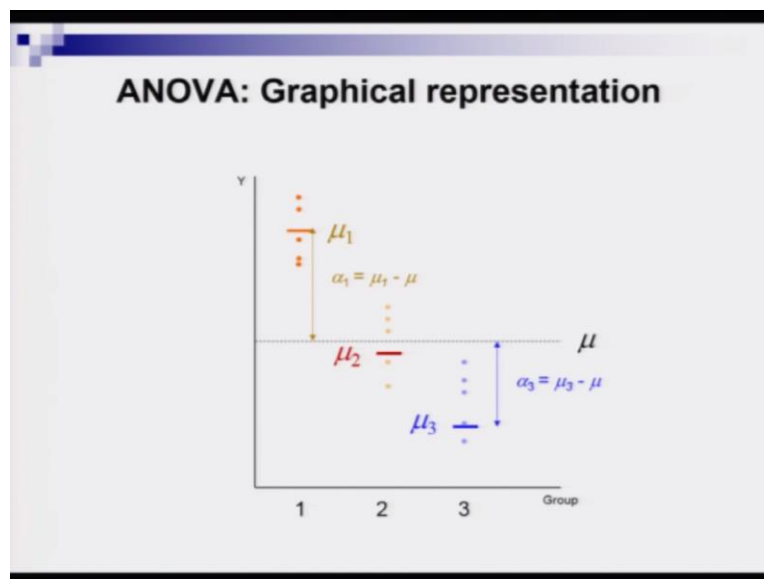
Now, what I am saying here by SSB, there is a dispersion between these sample means that you have computed. So, just for example, assume that you have seven samples and you are saying that, so these seven samples come from seven different populations, so  $k$  is equal to seven. So, for each of these seven samples, you can compute the arithmetic mean for  $y$ , your dependent

variable, continuous dependent variable. So, once they are computed, so like  $\bar{y}_1$  and  $\bar{y}_7$ , there will be a variation in these seven numbers. And that dispersion is captured by the SSB measure.

Now, what do I mean by the SSW measure? So, if you look at the second sub-bullet point of the last bullet point in the slide, I can define SSW as dispersion that exists among the data values within a particular factor level. So, note that here the key point is that it is within group, so it refers to the individuals per group.

So, now, let us come back to that example of  $k$  equal to 7. So, you have calculated the arithmetic means for seven groups, right,  $\bar{y}_1$  to  $\bar{y}_7$ . Now, note that if you fix a particular group, say,  $k$  equal to 5 or  $k$  equal to 3. Then there can be some observations in that group, and that is basically you can say for a corresponding level of the grouping factor that you are using, there are some number of observations. So, now, you can also calculate the dispersion of these  $y$  values which are under one particular level of one fixed factor or one grouping variable level. So, then the value of that dispersion is SSW.

(Refer Slide Time: 8:14)



So, now, let us look at one simple diagram and let us try to have a graphical depiction of these concept SSB and SSW and SST. And I have shown you another diagram in the previous lecture,

but let me have an even simpler diagram and see whether it helps you to understand the ANOVA identity better.

So, now concentrate in this 2d diagram. I am measuring the continuous dependent variable along the y-axis or the vertical axis, and I am plotting or showing different groups. Suppose I have three groups only, to have a simple story, and they are named 1, 2 and 3. Now, if you compute the arithmetic mean based on the 15 data points, then you will get an overall mean or grand mean, and that is denoted by this  $\mu$  here, big  $\mu$  in black colour. You see the broken line in the quadrant y and group, so that basically gives you the specific value for the grand mean or the overall mean.

Now, let us assume that we want to get a measure of the dispersion of these 15 data points. So, of course, we can have a measure like variance. So, if we do not divide the sum of square expression by the number of observations minus one, that degrees of freedom number, then we will have simply a sum of squared deviation from the arithmetic mean, not the grand or overall mean. So, that actually gives me SST. So, that is the total sum of squares.

Now let us concentrate what is happening in specific groups. So, let us focus on group number one first. So, you see there are five orange-colored dots that you figure out, and for these five observations, of course, you can compute the group specific mean. So, for these particular five observations you compute the arithmetic mean and you denote that arithmetic mean by  $\mu_1$ .

And note that by hypothetical construction, of course, there is no theoretical reason, this is just for illustration, there is a big gap between these group specific mean  $\mu_1$  and the grand mean,  $\mu$ . And that difference can be denoted by  $\alpha_1$ . So, if you remember that previous equation from the previous slide, that ANOVA equation, that  $\alpha_1$  is basically the group specific factor. It is a fixed factor and it is a systematic factor. So,  $\alpha_1$  actually is the fixed or systematic factor that is associated with the level 1 of the grouping variable that we are using in our analyses.

Similarly, one can find  $\alpha_2$  and  $\alpha_3$ , I am not explaining them, we have already had a lengthy discussion in the previous class. So, now, in this context, if I am interested in calculating

the SSB then how should I go about? So, in this context, if I want to compute SSB between group sum of squares, then how should I proceed? So, you see here you have three mu values. So, corresponding sample means can be calculated, right. So, then if you now calculate the dispersion between these three points, in fact, it is among three points, because it's more than two, then you get a measure for SSB.

Now, what is basically the SSW measure in this context? So, now, you have to focus on a particular group, say, group 1 or group 2 or group 3. In each group you have got five observations, so, now, you have to calculate that sum of square deviation from the group specific mean and that will be the SSW or sum of square within the group.

(Refer Slide Time: 12:33)

### ANOVA Identity

$$\sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{..})^2 = \sum_{i=1}^k \sum_{j=1}^{n_i} (\bar{y}_{i.} - \bar{y}_{..})^2 + \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i.})^2$$

↑

Total Sum of Squares (SST)

↑

Between Group Sum of Squares (SSB)

↑

Within Group Sum of Squares (SSW)

- ♣ This equation is the **ANOVA Identity**:
- ♣ To convert **Sums of Squares (SS)** into comparable measures of variance, divide the **SS** by their respective **degrees of freedom**
- ♣ Sum of squares (SS) divided by a *df* value, giving a **Mean Square (MS)**
  - $MSB = SSB / df_{between} = SSB / (k - 1)$
  - $MSW = SSW / df_{within} = SSW / (N - k)$

Okay. So, in the next slide, we are going to talk about the ANOVA identity once more. So, note that the slide starts with this big expression. So, this is an identity. And in the left hand side of this identity, we have what is called the total sum of squares, SST, and that can be partitioned or broken down or decomposed, whatever you want to call, in two components, namely SSB and SSW.

So, now, from this identity, we have to conduct a hypothesis testing to decide on this crucial question, whether the population group means are identical or not. Now, note that although we are talking about the population mean, we have started with the question regarding the

population mean, but the hypothesis testing that we are going to conduct that is going to talk about relationship between the variances, and we are going to show you why and how.

Basically, there is a theoretical reason behind it, because if you want to actually talk about comparison of population means across different groups, then you cannot make use of the T-test, I explained you why in the last lecture. So, what to do? So, then you can actually rewrite your model in such a way so that you can make use of a test that we have studied earlier and which has a better mathematical or statistical flexibilities compared to T-test, and that is the F test. And you will see that how we can go for a statistical hypothesis testing in this ANOVA context by making use of a F test, that we have learned before.

So, for F test, we have to first convert the sum of squares expressions into comparable measures of variance. Why? Because you remember that F is basically a ratio of two variances. So, you have to divide the sum of squares expressions by their respective degrees of freedom numbers.

So, how to proceed? So, this sum of squares, SS, say, has to be divided by the degrees of freedom value, and that will lead to a new measure and it is known as the mean square, which is abbreviated as MS. Now, there could be two possible MS values here because of this ANOVA identity, one will come from the first component in the right-hand side expression which is SSB, and the second one will come from the second component of the right-hand side of this identity, that is SSW.

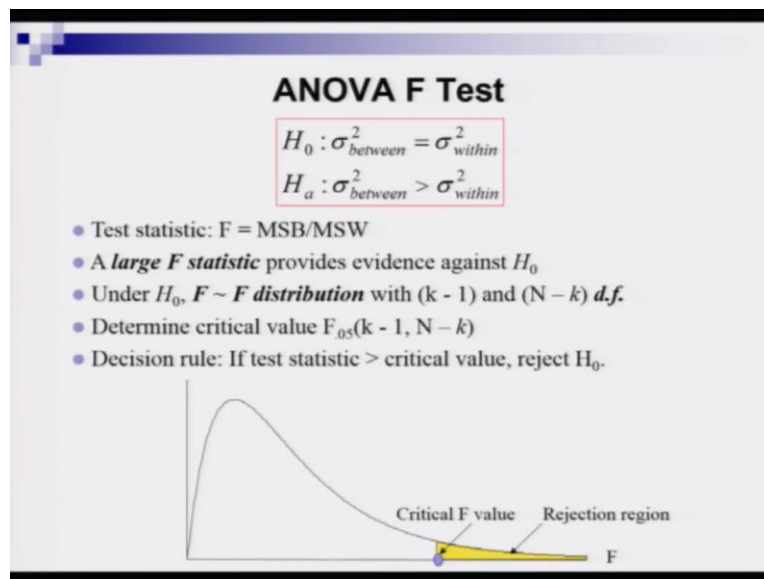
So, let us first talk about the mean square which is coming from the between group sum of squares. So, you have to divide SSB number that you have calculated from your raw data and you have to divide that by the degrees of freedom corresponding to SSB case. So, here, as we have k number of groups, the degrees of freedom is k minus 1.

Next, we move on to the mean square which arises from the within group sum of squares. So, MSW can be found by dividing SSW by the degrees of freedom that is relevant for the within group sum of square. So, here, we have total number of observations that is capital N, and you have to deduct the k number of sample group specific means that you need to compute. So, the degrees of freedom number will be capital N minus k.

So, in the next slide we are going to talk about the F test, which is associated with these ANOVA one-way process. So, here note that we are going to talk about a comparison of variances. And these variances are, to be very specific, these sample variances are coming from the ANOVA identity.

So, let us first start by looking at the null and alternative hypothesis, because we have started with a different kind of null and alternative hypothesis which was expressed in terms of the population means. Now, we are not testing the equality or inequality of population means, now our focus is on the variance part. So, we have to restate our null and alternative hypotheses in this one-way ANOVA context.

(Refer Slide Time: 17:22)



So, we start with the null hypothesis which says that, well, we can assume that the variance, which is related to the between group of sum of squares, that is equal to the sigma square or the variance that comes from the within groups calculation. And the alternative hypothesis, H-a, can be assumed that sigma square between is not equal to sigma square within.

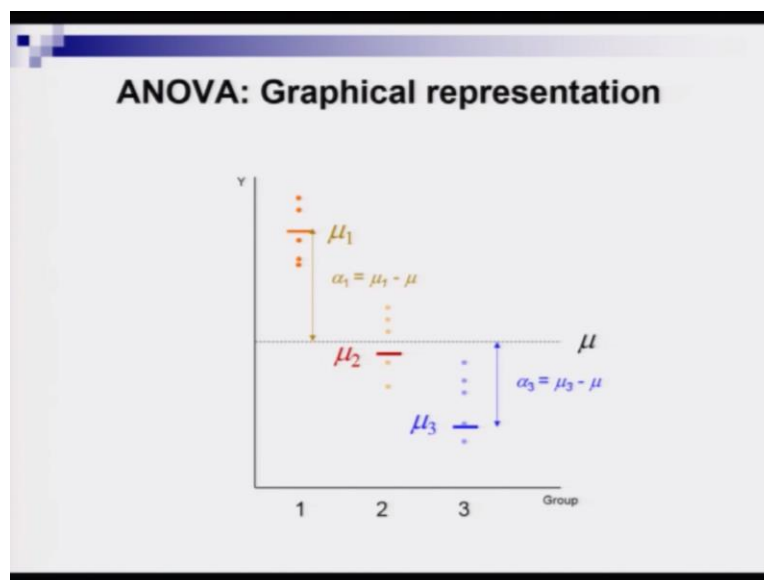
So, next we will look at the F statistic value. So, here, the test statistic is defined as capital F, and that is basically the ratio of MSB and MSW. Now, note that here I am writing MSB divided by MSW but not writing MSW divided by MSB. Why? Because, if you remember the previous lecture, there I have told you that the population mean will be significantly different for at least



one group if the between sum of squares values are actually larger than the within sum of square values.

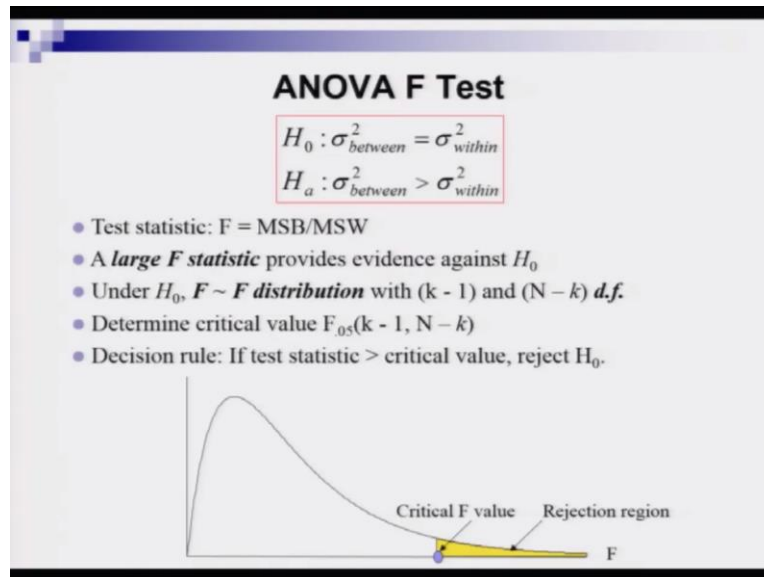
So, if the MSB is larger than MSW, then we can say that there could be a case for unequal population means. Actually, I said that you have to figure out which particular variance is higher and which particular variance is lower and then you keep that particular variance in the numerator which has got higher value. So, that is the way generally we proceed with F test. So, here, the diagram, let me take you back to the previous diagram.

(Refer Slide Time: 19:15)



Yeah. So, let us have a look at this diagram again. Here also you see that we have a strong evidence from the sample points that the between sum of squares has got the higher value than the within sum of squares, because the dispersion associated with the 15 points is higher than the dispersion that is associated with the individual five points within a particular group.

(Refer Slide Time: 19:51)



So now we are back to the old slide, and then let us consider the second point. So, the discussion on the first bullet point itself probably will be good enough for you to make a case that a large F statistic value actually provides enough evidence against the null hypothesis. Now, the question is, how large is large enough? So, for that, we have to conduct a proper F test by following the principles that we have learned before.

Now, if  $H_0$  is indeed true, then this F statistic, defined above, follows an F distribution with two degrees of freedom, one for the numerator and one for the denominator. So, here, the numerator is MSB, So the corresponding degrees of freedom is k minus 1, and the denominator is MSW, the corresponding degrees of freedom is capital N minus k, where capital N is the total sample size. So, that is basically the total number of observations in your data set. And let me remind you again, k is basically the number of groups.

So, then what you have to do, we have to set a level of significance. Again, you can start with alpha value equals 0.05 following the standard norm. So, then the next step would be to determine the critical value by consulting our statistical table and you get the critical value capital F 0.05 for two degrees of freedom, k minus one and n minus k. We already had discussion regarding how to consult an F table, so please consult the previous lecture.

Now, once the critical value is determined, we have to figure out a decision rule. And then as per that rule, you have to decide. Now, let us follow the traditional approach. And by following it, we can set up a decision rule. So, if our test statistic value is greater than the critical value that we got from statistical table, then we can reject our null hypothesis.

So, now let us have a simple diagram through which I can show you what is happening. Again, this is a recap, because we had a discussion on F test using diagram and all, but those who have forgotten that discussion, just I am going to spend one minute here to remind you about that.

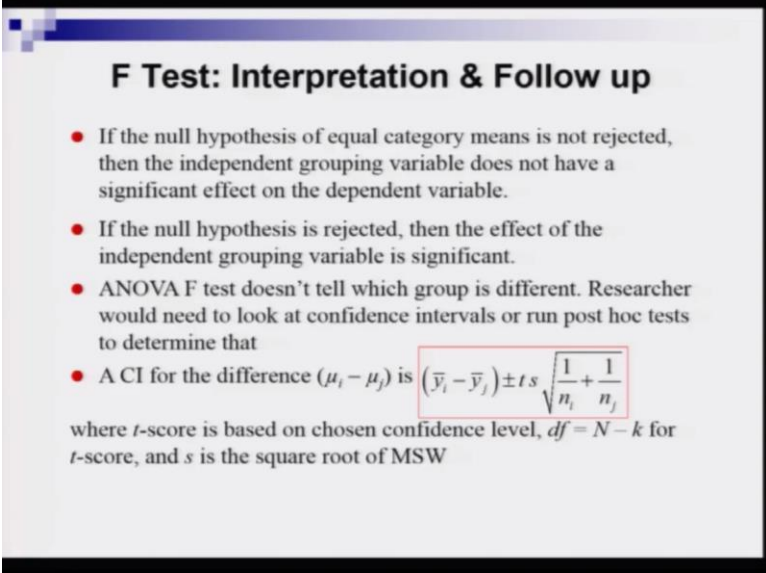
So, you remember that F actually is a positively skewed distribution, but its shape depends on the degrees of freedom values. So, let us assume that we know the degrees of freedom values, and for that this F actually takes this kind of shape that I am showing you here. So, F values are measured along the horizontal axis and along the vertical axis, we are measuring the probability values.

So, suppose from that table we got a critical F value, and here I have plotted that critical F value by drawing this blue circle. And you see that circle partitions this area under the F probability distribution curve in two parts, the one part is acceptance region and the other part is the critical region. So the area which lies at the right hand side of the curve, that is basically the rejection region and that is marked with yellow colour. So, here, we can talk about the decision rule through these diagrams in even simple way.

So, if you get a critical value which has higher value than this particular value shown by this blue circle, then that test statistic value actually falls in the rejection region and you should reject your null hypothesis.

So, this is the way you can conduct an ANOVA F test and you can take a call on the null hypothesis, whether to reject or not to reject the null hypothesis. But what next, shall we stop there? The answer is no. So, there are many interesting issues which are related to this F test and that we should have some knowledge about. So, in the next slide, I am going to summarize all these important points.

(Refer Slide Time: 24:23)



**F Test: Interpretation & Follow up**

- If the null hypothesis of equal category means is not rejected, then the independent grouping variable does not have a significant effect on the dependent variable.
- If the null hypothesis is rejected, then the effect of the independent grouping variable is significant.
- ANOVA F test doesn't tell which group is different. Researcher would need to look at confidence intervals or run post hoc tests to determine that
- A CI for the difference  $(\mu_i - \mu_j)$  is  $(\bar{y}_i - \bar{y}_j) \pm t s \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}$

where  $t$ -score is based on chosen confidence level,  $df = N - k$  for  $t$ -score, and  $s$  is the square root of MSW

So, now, we are going to have a brief discussion on the interpretation of the F test and the follow up exercises. So, let us first talk about how to interpret the F test. So, if the null hypothesis of equal category means is not rejected, or in other way round, if you cannot reject your null hypothesis then what should be the inference drawn from that result? So, in that case, you can infer that this independent grouping variable does not have any significant impact on the dependent variable  $y$ .

Now, what if the other way around? So, if the null hypothesis is indeed rejected, then what should you conclude? So, there you can draw an inference that the effect of the independent grouping variable is significant, because you see at least one of these sample group means is different from the overall sample mean. So, that is what your sample is showing. So, you can say that at least one population group mean is different from the other population group means.

Now, note that although we can make a qualitative statement that by rejecting, I can infer that at least one of the group means is not equal to the rest, you can't actually say which one is different from the rest. So, for that, you have to conduct some more statistical analysis, and next we are going to discuss about that very briefly.

So, if you reject your null hypothesis, then what next you do? So, you have to figure out or you have to get some idea that which population mean would be different from the rest. So, here, you

can actually look at the confidence intervals or you can run post hoc tests to determine this particular fact. But let us not talk about the post hoc test, because that is something we have not discussed. So, let us have the discussion regarding the confidence intervals only, because that we have discussed previously.

So, confidence interval, CI, for the difference between two population group means,  $\mu_i$  and  $\mu_j$  could be developed from whatever sample statistic we have calculated from the data. And there is a formula for that, and I am showing you the formula there in the red box. So, it's very simple to remember, so  $\mu_i$  minus  $\mu_j$  is now a random variable.

So, it is basically a population parameter which is unknown to us. So, if we assume that there is no difference between the  $i$ -th group population mean and the  $j$ -th group population mean, then we are assuming that the difference ideally should be 0. So, that is basically the null hypothesis. If null hypothesis is true then  $\mu_i$  is equal to  $\mu_j$ .

So, now, the sample analogue for this expression will be  $\bar{y}_i$  minus  $\bar{y}_j$ . So, these are basically the group specific sample means for the  $i$ -th and  $j$ -th groups, and you take the difference between them and that becomes a sample analog for that population expression.

Now, of course,  $\mu_i$  minus  $\mu_j$  is a random variable. So, for this population parameter, there will be a sampling distribution for the sample statistic  $\bar{y}_i$  minus  $\bar{y}_j$ . So, if there is a sampling distribution, there will be standard deviation and stuffs like that. So, basically this confidence interval is developed by either adding or subtracting this margin of error right.

So, how to compute that margin of error? So, for that we can actually go back to our T test, our good old friend students' T, because here you are just comparing two different population means, right. So, T is a perfect test in this case. So, here, the T score that you see in the formula is based on the chosen confidence level. So, you choose one minus alpha value, so it could be 0.95, it could be 0.99. So, based on that you can actually proceed.

Now, remember that T table actually gives you the values for small sample case. So, basically you have to look for or figure out the degrees of freedom as well. So, how to figure out the degrees of freedom in this case? So, here, in this case, you can write degrees of freedom equal to capital N minus small k, and that will lead you to the T score from the table.

And then there is another expression, small  $s$  in that formula box, what is that small  $s$ ? So, the small  $s$  is basically the square root of the MSW, so the mean square that emerges from the within group variation. So, we are done for the discussion on one-way ANOVA. And in the next lecture, we are going to start discussion on two-way ANOVA. So, come back for the next lecture. Thank you.