

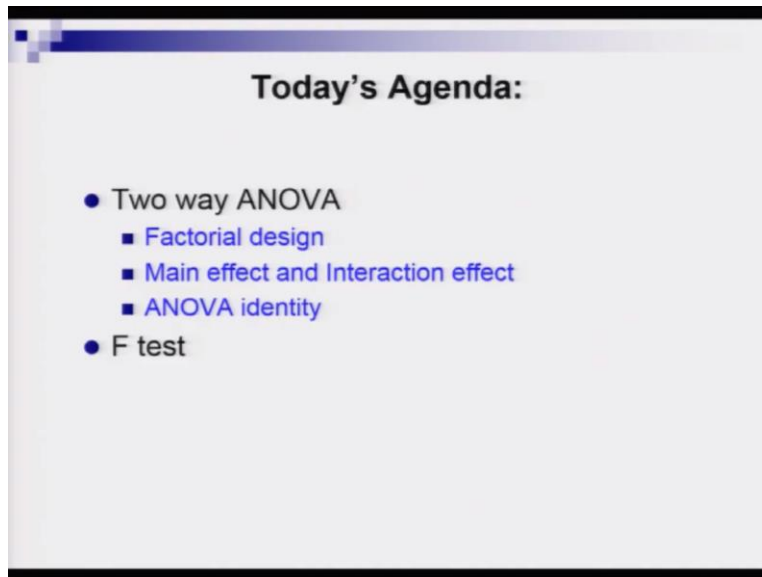
Applied Statistics and Econometrics
Professor Deep Mukherjee
Department of Economic Sciences
Indian Institute of Technology, Kanpur
Lecture-19
One-way ANOVA

Hello, friends, welcome back to the lecture series on Applied Statistics and Econometrics. So, in last two lectures, we have been discussing the case of analysis of variance. Today also we are going to continue the same discussion, but today we are going to study n-way ANOVA. Well, but, n-way ANOVA is a complicated matter, so that is why I decided to give you the case of two-way ANOVA which is simple to understand and easy to deal with.

So, what do we mean by two-way ANOVA? So, if you remember our discussion on analysis of variance so far, we have one continuous variable Y and we have one grouping variable, say, A , which is a qualitative variable. And this grouping variable A has two levels, say, a_0 and a_1 . And then as per this a_0 and a_1 , you group your Y observations under two different buckets, and then you are interested to know whether there is a difference in the population mean for Y across or between two groups or not.

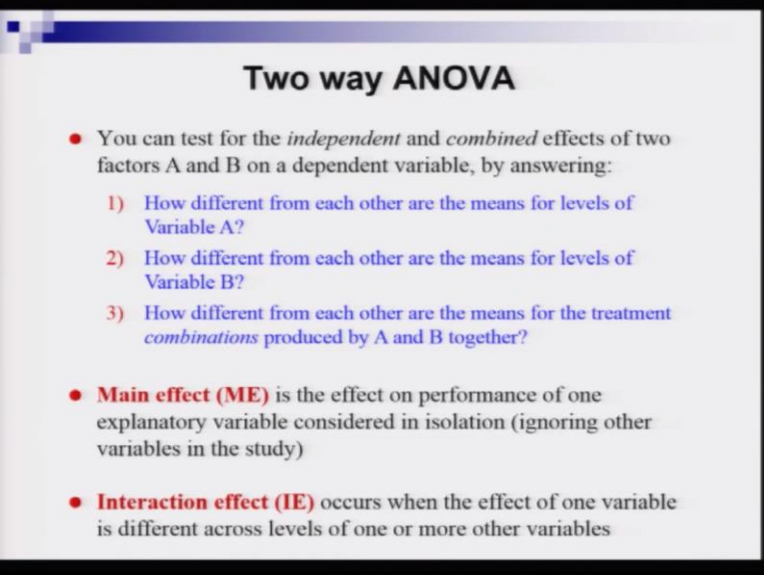
Now, what if, if I say that now you are interested to study the impact of another qualitative variable, say, B . So, then how would you proceed? So, that is going to be the subject matter of today's discussion, which is mostly based on that two-way ANOVA.

(Refer Slide Time: 1:50)



So, let us now have a look at today's agenda items. So, in today's lecture, we are going to briefly describe what is known as factorial design. And this factorial design is a very important concept or statistical tool in many fields, namely psychology, agricultural science, education research, medical sciences. And then we are going to study what is the difference between main effect and the interaction effect, and then we are going to revisit the ANOVA identity in the case of two-way ANOVA, and finally, we are going to show you how to conduct an F test in two-way ANOVA case.

(Refer Slide Time: 2:34)



Two way ANOVA

- You can test for the *independent* and *combined* effects of two factors A and B on a dependent variable, by answering:
 - 1) How different from each other are the means for levels of Variable A?
 - 2) How different from each other are the means for levels of Variable B?
 - 3) How different from each other are the means for the treatment *combinations* produced by A and B together?
- **Main effect (ME)** is the effect on performance of one explanatory variable considered in isolation (ignoring other variables in the study)
- **Interaction effect (IE)** occurs when the effect of one variable is different across levels of one or more other variables

So, we are going to start the discussion on two-way ANOVA by asking three questions. So, note that when you have two qualitative variables or factors A and B affecting a continuous dependent variable Y, then you can think about two different effects, and they are called the independent main effects and the combined interaction effect.

So, how do you get appropriate measures for these two effects? Because, if you can get the measures for independent main effects and combined interaction effects, then you can establish that your qualitative variables or factors A and B have some statistically significant impact on the dependent variable Y.

So, we start by asking three different questions. So, question number one is the following; how different from each other are the means four levels of variable A? So, what do we mean here? We mean that, okay, there is a continuous dependent variable, say, Y, and there is a factor A. So, let us assume that there are only two levels for this factor A.

So, if I now club my observations or if I now put my Y observations in these two different groups say, A0 and A1, whether the splitting of sample has any impact on the mean response or mean Y value or not. So, the same question could be asked in terms of the factor B, and that is what exactly I have written as the second question.

Now, the third and last question is even more important. When you are dealing with two or more explanatory variables, you cannot assume that these two independent variables are not going to affect Y simultaneously. There could be possibilities that two independent variables, in our case, two factors, A and B, could have an impact on Y simultaneously. So, there could be interaction between A and B. So, that will be the interaction effect.

So, if I want to know measure the interaction effect, I should get the answer for the question number three, which asks, how different from each other are the means of the response variable or outcome variable for different combinations produced by A and B together? So, the main effect is the effect measuring performance of one of the explanatory variable considered in isolation. So that means that we ignore the other factor or the variable in the study when we are changing the levels of this particular explanatory variable in question.

And what is interaction effect? So, interaction effect occurs when the effect of one variable is different across the levels of one or more other variables. So, basically, here, both the qualitative variables or the factors, A and B, their levels are simultaneously changing.

Now, I am going to talk about factorial design. So, I have told you at the very beginning of this lecture that factorial design is a very useful concept in applied research but it's quite complicated as well. So, for this course, I actually want to show you the simplest possible case of factorial design, and that is the two by two. So, let us have a look at a two cross two factorial design problem.

(Refer Slide Time: 6:33)

Factorial Design

- **Factorial design** is a statistical technique that allows for the estimation of the main and interaction effects between 2 or more independent variables on an outcome variable
- Example: Suppose the yield from different plots in an agricultural experiment depends upon fertilizer (factor A) and seed variety (factor B)

| | Fertilizer a_0 | Fertilizer a_1 |
|------------|------------------|------------------|
| Seed b_0 | I | II |
| Seed b_1 | III | IV |

- I vs. II: Effect of A at b_0
- III vs. IV: Effect of A at b_1
- If these are different, then we say that A and B interact
- I vs. III: Effect of B at a_0
- II vs. IV: Effect of B at a_1
- If these are different, then we say that A and B interact

So, let us begin by defining factorial design. What is it? So, it's a statistical technique that allows for the estimation of the main and interaction effects between two or more independent variables on an outcome variable. And here, the independent variables could be of qualitative in nature, so there could be attribute type variables and generally the outcome variable is a continuous one.

So, we will tell you the story of two plus two factorial design through a simple example. And this time I am going to pick the example from agricultural sciences. We all know that crop yield depends on the dosage of fertilizer or the type of fertilizer you are providing to the crops and also the variety of seeds, whether it's a local variety or whether it's an improved variety or it's a BT one. So, seed variety or seed quality definitely will have an impact on that crop yield, and so too the dose of fertilizer or type of fertilizer.

Suppose an agricultural scientist wants to measure the impact of different seed varieties or different dosage of fertilizer on crop yield and he wants to run an experiment, so how do you think that this person is going to proceed? So, here, let us assume that there are two factors, fertilizer, that is factor A; and factor number two is seed variety, that is my factor B. And both these factors, they have two levels. So, fertilizer factor A has two levels, a_0 and a_1 . And then seed variety factor B also has two levels, b_0 and b_1 .

So here, in the slide, you see I have a table where I am showing you for cells with no Roman I, II, III, IV written in these cells. So, what do they mean? So, they actually mean the combined effects of different levels of factors A and B. So, now, if I, for an example, look at the cell in the southeast, so that is basically number Roman number IV, you can see there, so what does it tell us? So, it tells us what is going to be the impact of combination of a-1 and b-1 on crop yield. So basically, if I apply the a-1 level of fertilizer and b-1 level of seed variety, then what kind of a mean crop yield I am going to expect?

So, of course, based on this simple setup, you can compare different cells, two at a time, and that is what we are going to do the next. And you see, I have listed down the results of these comparisons so that you can follow it quickly. So, let us look at the cells one and two. So, if I compare cells I and II, then what do I learn?

So, note that what is happening if you compare I and II, so you were fixing the level of seed where it at b naught and then you are changing the levels of fertilizer factor from a naught to a-1. So, basically what are you doing? You are measuring the effect of the quality factor or the attribute A at a specific value of the factor B. So, that specific value is of course b naught.

So, similarly, if you compare III and IV, then you get the effect of the attribute A for the specific value of attribute or factor B, and the specific value is b-1. So, now, if these are different numbers, then we can say that A and B actually, indirect. Now, similarly, you can compare the cells I and III, and II and IV, and then derive certain similar interpretations. So, now, we are going to discuss these two effects, main effects and interaction effects in detail in this context of this two cross two factorial design example.

(Refer Slide Time: 11:18)

2 × 2 Factorial Design: Effects

- In the absence of interactions, MEs have a straightforward interpretation: What happens to the mean outcome as we change the level of factor A and keep the level of factor B fixed?
- There is an interaction between A and B if the difference in means for the different levels of factor A changes as the level of factor B changes.
- If there are interactions, the main effects no longer have a clear interpretation.
- Find MEs and IEs is by getting **Marginal Means** which are the means for one level of a factor averaged across all level of the other factor.
 - ME of A: Is there a large difference among the A marginal means?
 - ME of B: Is there a large difference among the B marginal means?

| | A = a ₀ | A = a ₁ | Marginal Mean |
|--------------------|--------------------|--------------------|---------------|
| B = b ₀ | 10 | 15 | 12.5 |
| B = b ₁ | 15 | 20 | 17.5 |
| Marginal Mean | 12.5 | 17.5 | |

So, how to interpret ME and IE? So, in the absence of the interactions, the main effects have a straightforward interpretation, what happens to the mean outcome level? So, that is basically mean of my dependent variable Y, as we change the level of factor A and keep the level of factor B fixed. So, you can also change this phrase slightly, and you can switch the positions of A and B. So, if there is no interactions, you can get the main effects for both of your factors A and B. But if there is an interaction between A and B, then the interpretation becomes quite difficult.

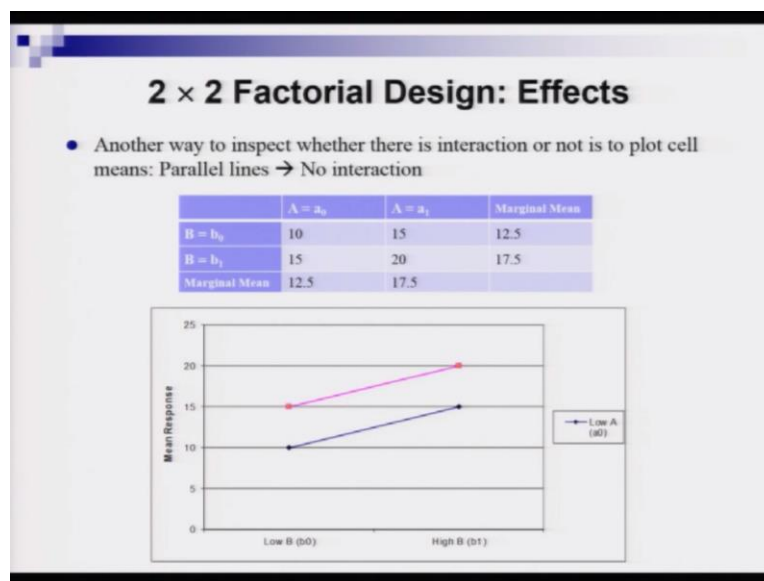
So, now the question emerges, how can I measure the main effects or how can I know from data that there is no interaction effect? So, for that, there are three different methods possible. So, first, we are going to look at the method of marginal means, then we are going to study the method of the diagram method, and the third one we are going to discuss the case of ANOVA method.

So, now, we are going to first define what do we mean by marginal means, because that is the pivotal concept for the method number one. So, marginal means are basically the means for one level of a factor averaged across all level of the other factor. So, this definition will be clear if we look at the table below. So, suppose I am interested to get the marginal mean of a particular value, say, a naught of the factor A.

So, then what I am going to do? So, I have to now fix the column here. So, the first column I am going to fix and then I am going to move down across rows. So, I see two different numbers for two different values of factor B, b_0 and b_1 , and they are 10 and 15. So, I have to basically now take a simple arithmetic mean of these two numbers and I get 12.5. So, that is basically the marginal mean for the attribute value a_0 for attribute A. Similarly, I can interpret the other marginal means that I am showing in this table.

Now, once the marginal means are computed, how can I say that these marginal effects are statistically significant? Well, we can't comment on the statistical significance of these marginal effects, you can maximum look at the difference between the marginal means and then see whether they are large enough or not. But the question remains, how large is large enough? So, how do you then make sure that you don't have any interaction effects? And if whatever you are observing the difference between the marginal means, are actually due to the main effects. So, for that you can draw a diagram, so that would be the second method.

(Refer Slide Time: 14:47)



So, here, in this slide I am going to talk about the second way of judging whether there is no interaction effect in your data. So, we will continue with the same old table that I have shown you in the previous slide. So, let us have a look at the case. So, here, at the bottom of the slide I am showing you a diagram here. So, on the y-axis or the vertical axis, I am plotting the mean response. So, you see, in the x-axis I am measuring two different values of the attribute or factor

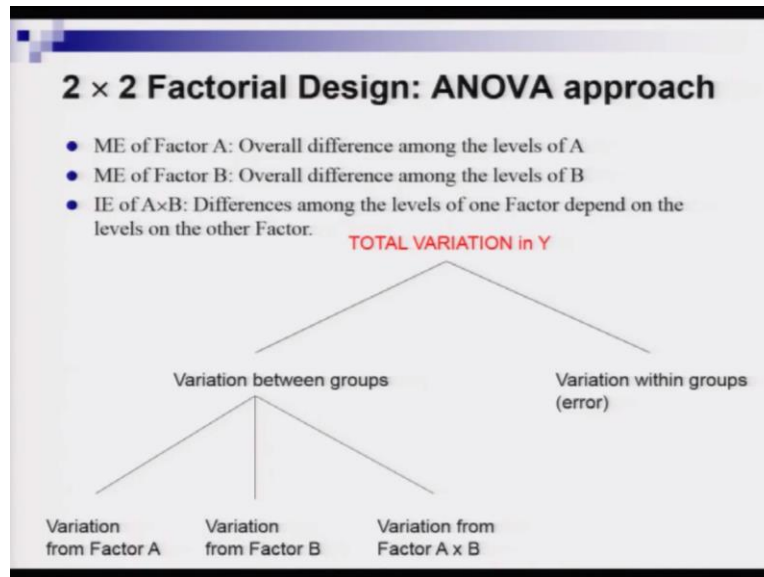
B, and I am coding them as low B, and that is the b naught value; and there is high B, so that is basically the b-1 value.

Now, I am going to fix low B, b naught category, and for that I am going to now plot the values or the cell means for capital A equal to a naught and capital A equal to a-1. So, for capital A equal to a naught, the value will be 10, and that is what you are seeing here, these black colour or blue colour diamond sign. And then the second value would be 15, why? Because I am fixing the value of b naught and then the other case possible is capital A equal to small a-1? And my cell means suggests 15 for that particular combination b naught a-1. And that is what I have plotted.

So, you see 15, that is basically square with red colour border and red box that you are seeing. So, here, similarly you can draw the other two points or the other two cell means, and note that if I join these cell means meaningfully, then I get two parallel straight lines, positively slope. So, here, by following the statistical theory, I can tell that there is no interaction, but, of course, this is a very simplified assumption and this is just an illustration, there could be interaction between A and B.

And if interaction is there, then there could be several types of diagram can emerge, there could be like the one line crossing the other one, or if it is not crossing the other one the slope may differ from one to the other. So, there are many possibilities if there is indeed an interaction between two qualitative factors, A and B. But you see, if you get to parallel straight lines, that is a simple rule. So, if you observe after plotting the cell means that you can obtain two parallel straight lines, then that means that these two factors, A and B, they are independent to each other. So, there is no interaction happening.

(Refer Slide Time: 18:00)



So, now, I am going to talk about the third approach, which is the ANOVA approach. Now, why do we require a third approach, because we have gone through two different approaches and they seem to be pretty easygoing, so they are simple. But you understand that real life is very complicated and simple methods may not be able to handle all the complications that we observe in reality.

So, if you have, say, more than two grouping variables and, say, two levels per grouping variable, then what will happen, can you draw a diagram and then see whether you get two parallel lines or not? Probably not. And also, you note that when you do the math by looking at the marginal means, and the diagram that you draw, they may actually give you contradictory results.

So, in that case, how do you actually conclude whether there is interaction effect or not? Or whether the main effect or the interaction effects are significant or not? Note that first two approaches that I have shown you here, they are kind of summary measures, right, you cannot establish statistical significance. So, here, I am going to talk about the ANOVA approach, which is useful when you have a complicated story to deal with or you have some contradictory results emerging from first two approaches.

So, note the difference or the similarity between the previous two lectures on ANOVA and their today's lecture. So, if you remember, I started with the ANOVA story in the very first lecture by comparing the population means. And later, I said that, well, it is better if we look at the variance and that is what gives rise to this ANOVA identity. And finally, we ended up conducting an F test which is a test for variance.

So, here also if you see, the first two approaches that I told you, which are extremely simple, the cell mean plotting and marginal means approach, they are also somehow comparing different mean values, and they are not talking about the variance. But here, the third approach, the ANOVA that we are going to study next, is going to apply the concept of variance because it is a general one and we can actually talk about statistical testing easily if we actually apply this concept of ANOVA in the case of two by two factorial design models.

So, here, in a simple two cross two factorial design world, how do I measure the main effect of factor A? So, basically, I have to focus on the overall difference among the levels of the factor A. And similarly, we can also get an idea about the main effect of factor B by looking at the overall difference among the levels of factor B. And how do I get the measure for interaction effect? So, we are talking about A cross B. So, the differences among the levels of one factor will depend on the levels on the other factor here. So, that is the interaction effect in nutshell.

Now, let us see how my ANOVA will look like in this case. So, I am interested in modelling the total variation in Y, and that can be broken down in two parts, if you remember then ANOVA entity from the previous lectures. One is basically the variation between groups, and that is measured or abbreviated as SSB, and the other component is variation within groups or error, so that is abbreviated as SSW or SSE.

Now, the first component, variation between groups or SSB, can be broken down into three sub components here. And the first one will come from the variation from factor A, the second component will come from the variation from factor B, and the third component is the variation that will come from the interaction effect if it is significant, and that basically is a new factor that one can think about A cross B.

(Refer Slide Time: 22:50)

Two way ANOVA: Data Structure

| Factor A | Factor B | | | |
|----------|-----------|-----------|-----|-----------|
| | 1 | 2 | ... | b |
| 1 | y_{111} | y_{121} | ... | y_{1b1} |
| | y_{11n} | y_{12n} | ... | y_{1bn} |
| 2 | y_{211} | y_{221} | ... | y_{2b1} |
| | y_{21n} | y_{22n} | ... | y_{2bn} |
| : | : | : | ... | : |
| a | y_{a11} | y_{a21} | ... | y_{ab1} |
| | y_{a1n} | y_{a2n} | ... | y_{abn} |

Observation k
in each cell
↓
 y_{ijk}
Level i Level j
Factor A Factor B
 $i = 1, \dots, a$
 $j = 1, \dots, b$
 $k = 1, \dots, n$

Now, we are going to look at the data layout for a two-way ANOVA model. So, here, let us assume that the continuous variable is Y and we have two factors, A and B, and capital A has a levels and factor B has b number of levels. So, you see here in this slide, I measure or I denote an observation k by y_{ijk} .

So, here i is basically the levels for multi factor capital A, and there could be small a number of levels available for the factor capital A. The second component in the subscript is j, and that denotes basically the levels of factor capital B, and there could be small b number of levels available for the factor capital B. And then the third symbol in the subscript, that is k, so that is basically number of observations in each cell. So, here, I am assuming a simple case, I am assuming that there are n number of observations. So, k will take value from 1 to small n.

So, now, let us have a look at the matrix design that I am showing you here. So, in the rows I am denoting or I am representing different levels of factor A, and you see there are small a number of rows. And then in the columns, I have small b number of columns, so this are for each levels of factor B. So, what do we observe in the cell then?

So, say, one particular observation let me concentrate, and that is y_{121} . So, what does that mean? So, it means that, this is the observation that corresponds to the level 1 for factor 1, level 2 of factor 2. And this is the very first observation in the combination, and there could be n number

of such observations which are satisfying that they actually are corresponding to the level 1 of factor A and level 2 or factor B. So, this is the way you can interpret all other numbers or symbols that you are seeing in this data matrix.

(Refer Slide Time: 25:43)

Two way ANOVA Model

- The observed response is given by: $Y_{ijk} = \mu + \alpha_i + \beta_j + \epsilon_{ijk}$
 - Y_{ijk} = Variable value
 - α_i = effect of being in group number i
 - β_j = effect of being in group number j
 - ϵ_{ijk} = idiosyncratic error $\sim N(0, \sigma^2)$
- Need to estimate all μ and error variance σ^2
- Total variation in Y or Total Sum of Squares (SS_{Total}) can be split using ANOVA identity: $SS_{Total} = SS_A + SS_B + SS_{A \times B} + SS_{Within}$

| Source of variation | Degrees of Freedom | SS | MS | F |
|---------------------|--------------------|-------------------|-------------------|---------------------------------|
| A | a-1 | SS_A | MS_A | MS_A / MS_{within} |
| B | b-1 | SS_B | MS_B | MS_B / MS_{within} |
| A X B | (a-1)(b-1) | $SS_{A \times B}$ | $MS_{A \times B}$ | $MS_{A \times B} / MS_{within}$ |
| Within | ab(n-1) | SS_{within} | MS_{within} | |
| Total | abn-1 | SST | | |

Now, we are going to revisit the equation, the mother equation for the ANOVA analysis. And I am going to now show you how a particular observation y_{ijk} can be broken down in various components. So, here, you see the observed response y_{ijk} can be broken down into four components, one is the mean, which is the grand mean or the overall mean in the data of the response variable Y . And then there is α_i , so that is basically the effect of being in group number i . And now this new addition you see here, that is β_j , so that is the effect of being in group number j . And then I of course have the idiosyncratic error, ϵ_{ijk} .

So, needless to say that we need to estimate all these unknown population parameters, μ 's and the error variance σ^2 . Now, how to go about it? We all know that we have to make use of the ANOVA identity. So, in this case, what will be the form of ANOVA identity? I am not going to show you a complicated messy expression here involving some notations and all, so I am going to show you here a simple expression here and that is easy to follow. In the next lecture, I am going to give you an empirical illustration or an example so you know how to compute these components. But as of now, you focus on the ANOVA identity.

So, the sum of squares total, SS-total can be broken down in four parts here. So, one sum of squares will come from the factor A, the other one will come from factor B, and there is another one will come from the interaction effect which is A cross B. And finally, there will be one sum of squares coming from the errors, so that is SS-within. When we are conducting ANOVA, there is a concept called ANOVA table. In the previous two lectures I have not shown you the form it takes, but let us now have a look at ANOVA table. Because in different textbooks or in project reports, you are going to face this ANOVA table.

So, now we focus on the bottom part of the slide, here I am showing you the ANOVA table. So, you see there are five columns here, and these five columns you will always see by the way in any ANOVA table. So, the first column will give you the sources of variation. And here, of course, we have four sources of variation coming from the ANOVA identity. And the second column gives the degrees of freedom, I explained you why we have to take care of the degrees of freedom so I am skipping the discussion here, you can go back to the previous two lectures.

Then the third column talks about the numbers that you calculate from your sample and these are the sum of square numbers that will come from four different sources of variation in this particular case. So, if you divide the sum of squares by the degrees of freedom, then you get the means square, and that basically again can be computed for four sources of variation in this context. So, the fifth column gives me the F values.

So, how do you calculate the F value? So, you basically have to focus on a particular row and you get the MS value for that particular source of variation, and if you divide that number by the MS within number, then you get the F value corresponding to the source of variation. And of course, you will get three such F values, one for A, one for B and the other one is for a cross b. So, this is basically the structure of the ANOVA table in a nutshell.

So, now after ANOVA table, what to do next/ So, next we will conduct an F test, because we have already calculated the values of the F statistics. Now, see that we are mostly interested to figure out the main effects that it's not a bad idea to have a test for the interaction effects as well. Because, of course, we may be interested to know whether there is interaction effect, so whether A and B jointly have an impact on Y. So, we are going to conduct three F tests, two for two main

effects, one for factor A, one for factor B, and the other one is basically an F test which will test the statistical significance for the interaction effect, A cross B.

(Refer Slide Time: 30:17)

F Test for Two way ANOVA

- Test for the difference between the levels of factors A and B:
$$F = \frac{MSA}{MSW}$$

Rejection region: $F > F_{\alpha, a-1, n-ab}$
- Test for the difference between the levels of factors B and A:
$$F = \frac{MSB}{MSW}$$

Rejection region: $F > F_{\alpha, b-1, n-ab}$
- Test for the interaction between the levels of factors A and B:
$$F = \frac{MS(A \times B)}{MSE}$$

Rejection region: $F > F_{\alpha, (a-1)(b-1), n-ab}$

So, here, I am showing you how you can conduct this F test. We have done F test in the previous lecture only, so probably it is fresh in your memory. So, in the previous slide I have already shown you how to calculate the F value for the sources of variation. So, here, I am showing you three different cases.

So, the first one in the northwest corner of the slide is basically talking about the case of factor A. So, for that, I am showing you the rejection region. So, you have to calculate the value of F and then you have to figure out the critical value from the F table for a corresponding value of alpha, and the degrees of freedom. And if that calculated F value is higher than the critical value or the tabulated value, then you say that, okay, I reject my null hypothesis.

So, similarly you can follow the procedure to test for other F value. So, in the northeast corner of the slide, I am showing the case for factor B. And under the second bullet point of the slide, I am showing you the case for the interaction effect. So, we are done for today. In the next lecture, I am going to show you a numerical example, through which it will be easier to understand the calculation steps. And I would finish our discussion on analysis of variance by introducing a concept called ANCOVA. See you then. Thank you.