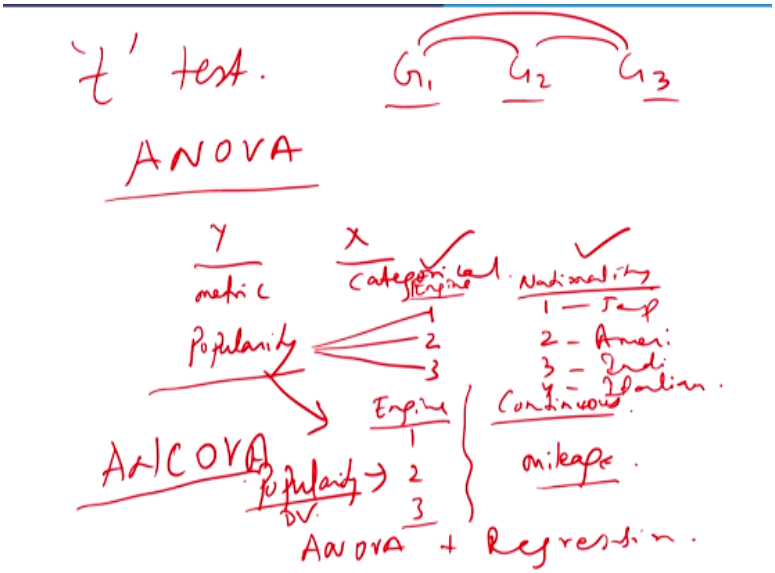
Marketing Research and Analysis-II
(Application Oriented)
Prof. Jogendra Kumar Nayak
Department of Management Studies
Indian Institute of Technology – Roorkee

Lecture – 35
Solving n-way ANOVA - I

Hi friends I welcome you all to the course Marketing Research and Analysis. So in the last class, we discussed about the analysis of variance.

(Refer Slide Time: 00:40)



So analysis of variance is a technique which is an extension of the t test, extension of the t test where you were earlier able to compare between 2 groups. So when a necessity occurred or arised for comparing more than 2 groups or 3 levels, for example 3 levels or 3 groups, then such a condition a multiple t test seemed reasonably correct but that had its own problems. The problem was that a multiple t test as I have explained earlier leads to an inflated alpha.

That means the type I error gets inflated because each time you take a 5% level of significance or whatever levels of significance that gets added up as you conduct in multiple number of t tests, that means for example as I had said earlier also, there are 3 groups. So if you have 3 groups, then multiple tests you will be doing t test so that means what 3 tests when you do each time 5% level of significance, the type I error somewhere goes and increases.

So to avoid that we had one test called the analysis of variance. This test is a very powerful test and highly utilized in experimental studies. So what is this test says is that whenever there are more than 3 groups or levels in an independent variable and the dependent variable is metric in nature or continuous in nature, in such a condition that means Y and X, Y is dependent which is my let us say metric I am saying and X is my categorical variable.

So let us say for example I am trying to say the popularity of a car is dependent on the type of engine it is using. So suppose there are 3 types of engines which is using, now this type of engine is leading to the popularity of the car, now that is what you want to suppose test. In such a condition, we would say this is a case of an analysis of variance. However, if there is only one factor you said there is for example is said here type of engine.

Suppose I would say the nationality, which country it comes from, for example let us say again we have 2 or 3 or 4 whatever, we have 4 levels. Now for example Japanese, American, let us say Indian, and the fourth is Italian let say. So suppose I am having 4 nations from which the car comes and I feel the popularity of a car is not only related to the type of engine, but also from the other nationality that it comes from.

In such a condition, we have 2 factors, one factor the engine and the other factor the nationality. So we have 2 factors, so we say it is n-way ANOVA and two-way ANOVA. A 2, 3, 4 all together we say they are called n-way ANOVA basically. So in the last class you also had discussed in the same condition if you have a case where the dependent variable is let us say same popularity and the independent variables is one is engine, let us say engine which is you measured as 1, 2, 3.

Other is a continuous variable, another continuous variable which is your independent variable. Now let us say what can it be? Let us think. So popularity is affected by the type of engine and let us say the mileage the car shows also leads to a popularity. Now mileage is measured in a continuous variable. Now these 2 things, you see type of engine and the mileage, they also could have a relationship also.

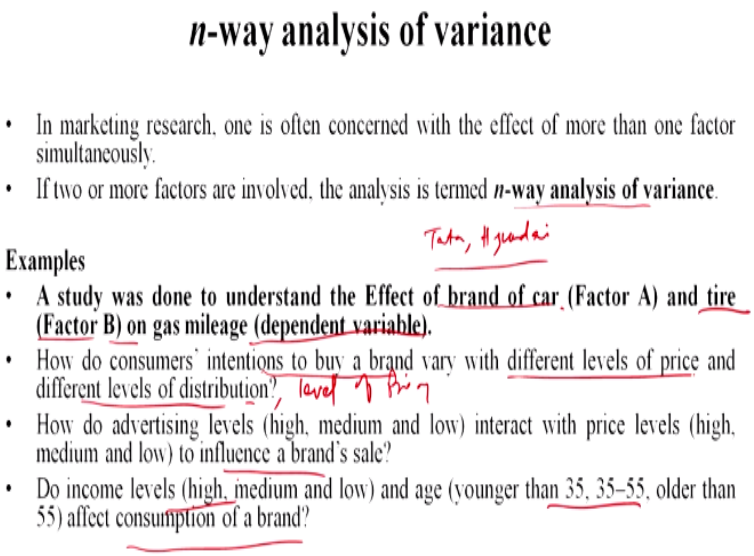
So in this 2, when I am having one independent variable as a categorical and other independent variable is a covariate which is a basically continuous variable, then a case of ANCOVA is used. The ANCOVA is nothing but a combination of ANOVA and a regression

model, so it is a ANOVA plus regression. As you can understand, you can see here so popularity is my let us say dependent variable which is continuous and engine is categorical.

So if you take up to this much, then it becomes a case of a ANOVA. When I am adding the dependent variable is metric and another independent variable is metric, then in that case what it becomes, it becomes a case of regression, so that is why I said ANCOVA is a case of ANOVA plus regression. So today what we will do is we will try to solve some problems by hand, since in the last class we had solved a problem by hand for ANOVA, today we will go little further.

Now what is this analysis is of variance? Let us understand in marketing condition.

(Refer Slide Time: 05:47)



In marketing research, one is often concerned with the effect of more than one factor simultaneously. If 2 or more factors are involved, the analysis is termed as n-way analysis of variance. Take some examples to understand, let me first clear what n-way analysis of variance means. A study was done to understand the effect of brand of car and tire on gas mileage. Now what is it saying?

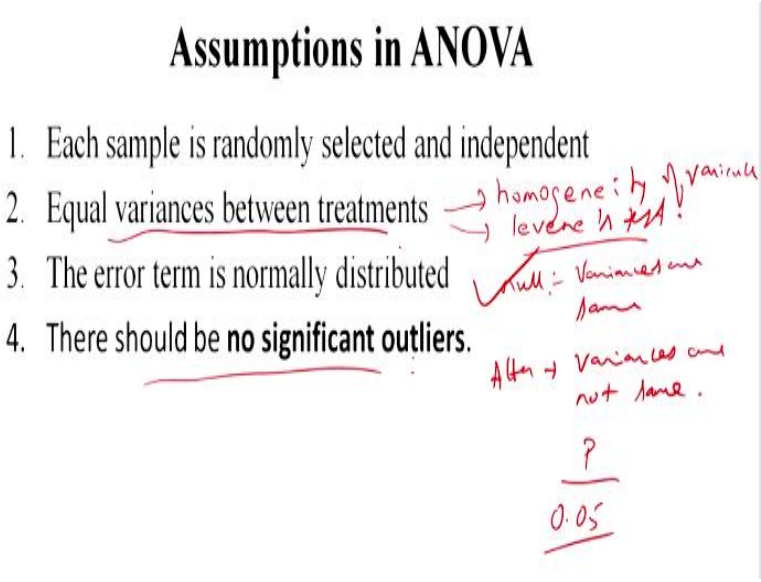
The brand of the car, which brand, for example Tata in India for example or the Hyundai car or what brands are you using and the tires that you are using how they affect the mileage. So mileage is my dependent variable and the brand of car and the tire which are both categorical in nature are my independent variables. Another example how do consumers intensions to

purchase a brand or buy a brand vary with different levels of price and different levels of distribution?

Now my intention to buy a brand is my dependent variable. My independent variables are levels of price and levels of distribution. So there again 2 conditions here, 2 cases. I could have used a third also, may be levels of price, levels of distribution, and level of promotion or in which media they are promoting let us say promotional medium TV 1, let us say radio 2, or print 3. So level of promotion I am saying. So I may add 3 also or more n.

Another case how do advertising levels interact with price levels to influence a brand sale, something close to this one. Last one does income level high, medium, and low and age affect the consumption of a brand. Now affect consumption of a brand is my dependent variable and these 2 are my independent variables. So you see in these conditions, we always have a dependent variable which is let us say continuous and independent variables are categorical.

(Refer Slide Time: 07:59)



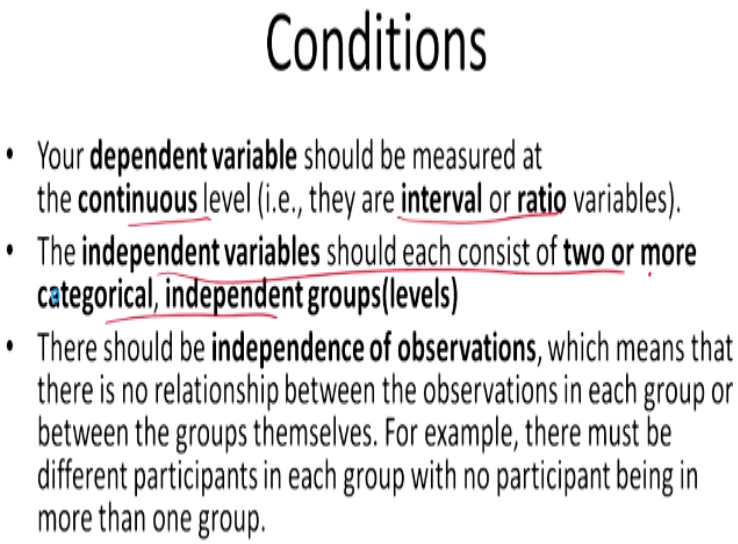
Now these are the assumptions in ANOVA. Each sample is randomly selected and independent. So these samples are independent of each other. They should not be, it is not a repetition. So if you want to do a repetition, then that is a special kind of ANOVA called repeated measures ANOVA, but here we are talking about that the samples or respondents are independent they will fall into one and only group.

Second equal variances between treatments, now this is a very important assumption, we call homogeneity of variance. Now it says it means that the variances between the different groups which you are trying to compare should be same, else they should not be compared. This is very popularly told as Levene's test which helps to measure the homogeneity of variance and it is very interesting to know that this Levene's test when you measure you have to have your null hypothesis is that the variances among the groups are same.

And alternate is that variances are not same, variances are not same, at least one group is different. Now surprisingly in this case since we are saying equal variances between treatments we wanted so we want our null hypothesis to be accepted, that means if you want to check your p-value, your p-value should be less than 0.05, suppose at a 5% level of significance you are checking, it should be less.

Generally what we do, we expect that we always reject the null hypothesis and accept the alternatives, but in this condition this special case, we want that we will accept the null and reject the alternate that is the condition. The error term is normally distributed right. So there is a normal distribution of the within variance or the error term. No outliers, significantly big outlier, because small outlier should be there.

(Refer Slide Time: 10:15)



What conditions are required? Dependent variable should be measured at the continuous level, so there is interval or ratio. Independent variable should consist of 2 or more categorical independent levels. Now there should be independence of observations which I just explain, which means that no relationship between the observations in each group or

between the groups themselves. For example there must be different participants in each group with no participant being in more than one group.

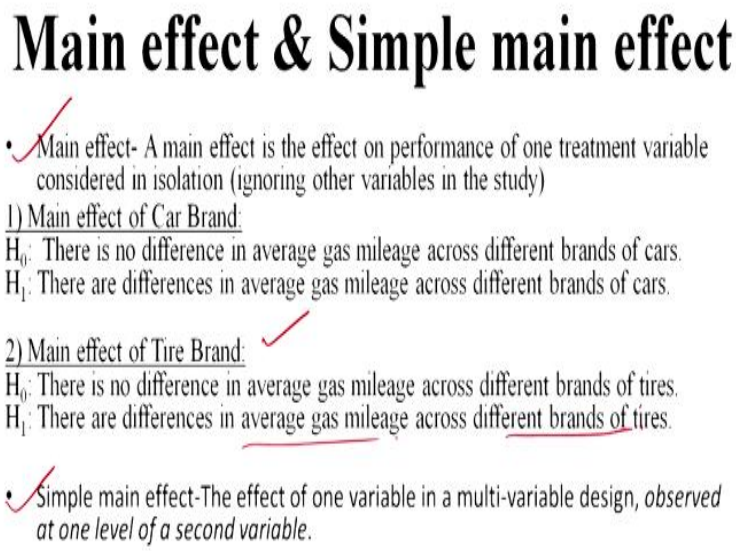
That means you cannot repeat the participant, so that is one of the conditions that you need to check. Now the same case if you see go back to this one, the first case. So I have tried to draw a small table.

(Refer Slide Time: 10:55)

	MRF	Apollo	Goodyear	Dunlop
Swift	10	11	9	8
Zen	12	10	10	9
Nano	14	12	12	10

So there are 3 brands of Indian cars; Maruti Swift, Zen Estilo and this is Tata Nano. There are 4 brands of let us say tires; MRF, Apollo, Goodyear and Dunlop. The mileages are these are the mileages which are given to you. So now in such a condition, we want to see the effect of not only the tires and also the brands together on the mileage of the car. So now when we talk about the n-way analysis of variance or two-way or three-way or four-way analysis of variance, we need to understand some terms, what is that?

(Refer Slide Time: 11:39)



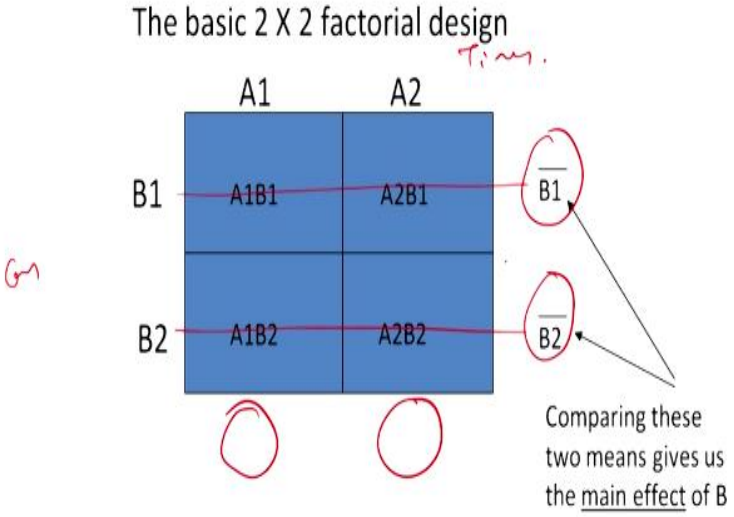
Now the first is the main effect. So when you are studying the effect of ANOVA, there are 3 effects basically. One the main effect, the second called the simple main effect, the third is called the interaction effect. First let me explain to you one by one very clearly what it means. First main effect, a main effect is the effect on performance of one treatment variable considered in isolation, that means what ignoring other variables in the study.

Suppose there are 2 variables, then you want to see, let the variables are A and B, you want to see the effect of A ignoring B, similarly you want to see the effect of B ignoring A, it is called that we are checking the main effect of A and B. Now in the last example in this the main effect of car brand case, you want to see the main effect of car brand, the brand of the car. What is the null hypothesis, there is no difference in average gas mileage across the different brands of cars.

So this is my null hypothesis, but I do not want to see that. So what is my alternate, there are differences in average gas mileage across the different brands of cars or at least in one of these cases, at least there is some difference, in the 3 cars at least one of them is different from the rest. Similarly what is the main effect of the tire brand, you try to write yourself first before seeing this.

So null hypothesis is there is no difference in average gas mileage across the different brands of tires and alternately the alternate hypothesis is there is at least some difference in the average gas mileage across the different brands of tires.

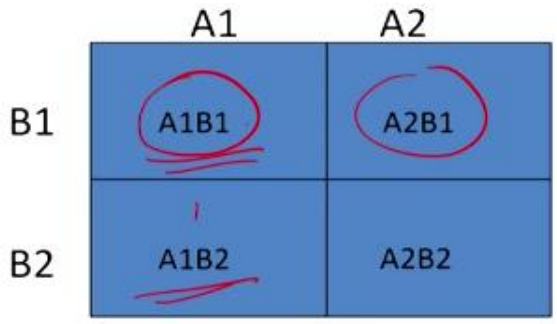
(Refer Slide Time: 13:27)



See this. So this is a case suppose we are taking this a 2 cross simple matrix. So let us say these are my tires and these are my brands of cars. Now what am I saying. The effect of A, This is A1, this is A2. So when I compare the effect, compare between these 2 means that is of A1 and A2, it gives me the main effect of A, that means if there is a significant difference between A1 and A2, I can say that the main effect of A is significant, that means A has a significant effect.

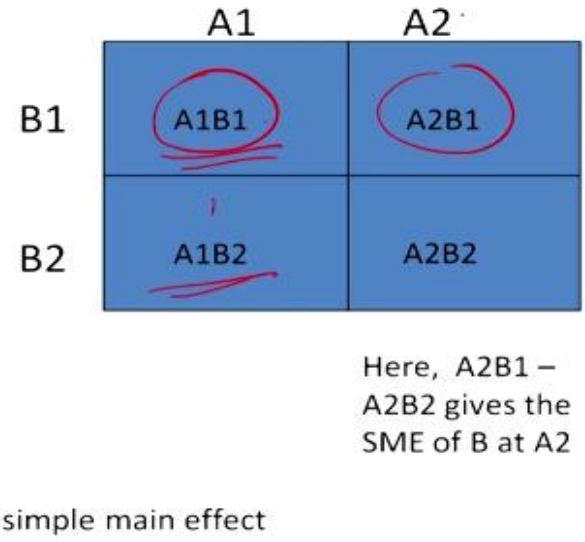
Similarly if I take B, so this is the side B, so if I take the difference between B1 and B2 and I see whether there is significant difference or not. If there is a significant difference, I would say the main effect of B is significant, otherwise I would say it is insignificant. Now coming to the third term is the simple main effect, which first let me show you what it means. What it says the simple main effect is the effect of one variable in a multivariable design, in our case the 2 x 2 matrix observed at one level of a second variable. Let me show you what it means, let us see this.

(Refer Slide Time: 14:50)



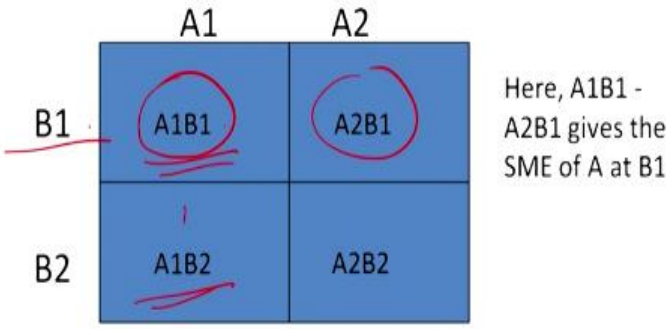
Now coming to this you see there is A1, A2, B1, B2. Now A1B1 is the level of A1 at B1, A2B1 is the level of A2, the value of A1 at B1. Similarly these 4 right. So here A1B1 minus you see this one, these 2, minus A1B2, this value this and this gives the simple main effect of B, why B is the common, of B at A1, now understand.

(Refer Slide Time: 15:38)



So now what you do is you write the simple main effect of B at A2 similarly. Let us see. So A2B1-A2B2 gives the simple main effect of B at where A2, so at what level it is playing. So as you go back to the refreshing set again, what is it saying the effect of one variable in a multivariable design observed at one level of a second variable. So the two variables are A and B and accordingly you are checking the simple and the main effect.

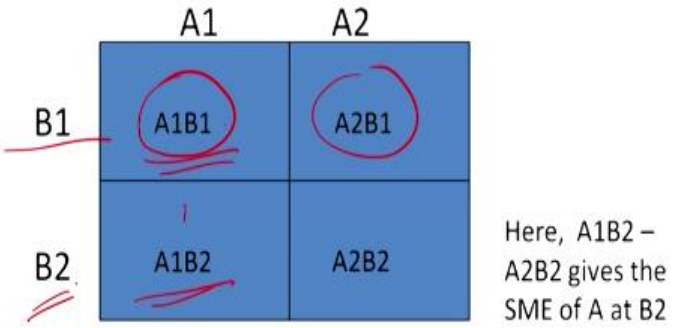
(Refer Slide Time: 16:21)



SME = simple main effect

Similarly this gives me the A1B1-A2B1 gives the simple main effect of A, now this is A at where in B1.

(Refer Slide Time: 16:26)



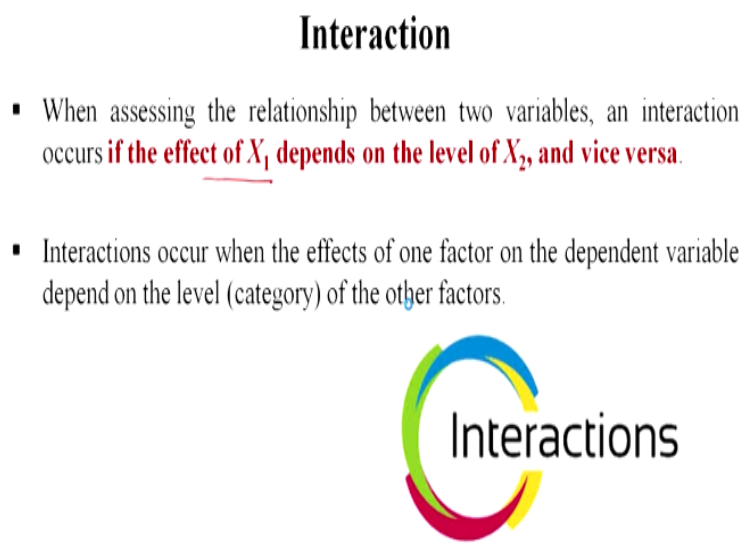
SME = simple main effect

Similarly A1B2-A2B2 gives the simple main effect of A at where B2. So these 4 things are to be are called as simple main effects. Now coming to another very important term which what happens in real life is that sometimes 2 variables would be there and individually they might not effect, but when they come together they play a very significant effect. For example we know that sodium and water, when sodium alone is there it does nothing and water alone does nothing.

But when sodium and water come together, they react and it gives a very serious reaction makes a very serious reaction. Similarly there are lots of things. Petrol itself will not do

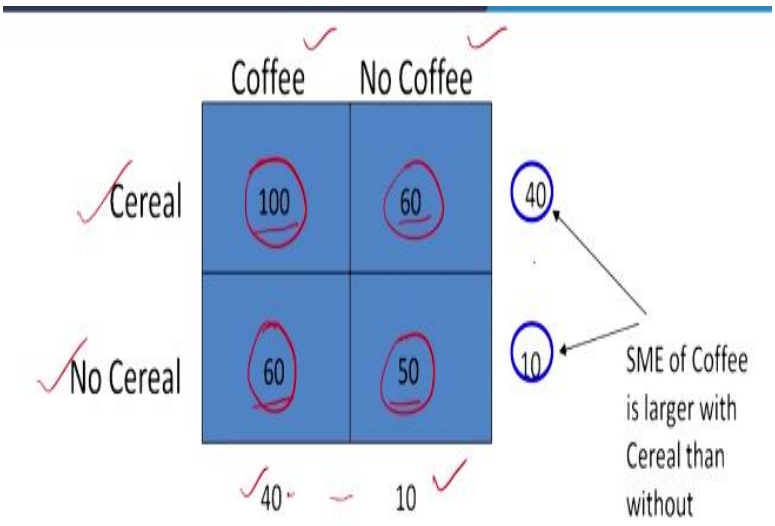
anything and let us say you put some chemical in that for example which can catch fire, so that when it comes close to petrol or any combustible substance, it gives rise to fire which can be dangerous. So you have to understand that there are certain things in life and they have a profound effect, which individually they will not do any wrong or any good but when they come together there is an interaction effect and the interaction effect plays a very very vital role. Now what is this interaction? Let us see.

(Refer Slide Time: 17:46)



When assessing the relationship between 2 variables an interaction occurs if the effect of X1 depends on the level of X2 and vice versa that means the effect of X2 depends on the level of X1, so at what level X1 is? If the value of X2 depends on that, then we say there is an interaction effect. Interactions occur when the effects of one factor on the dependent variable depend on the level of the other factors. So let us see this example.

(Refer Slide Time: 18:23)



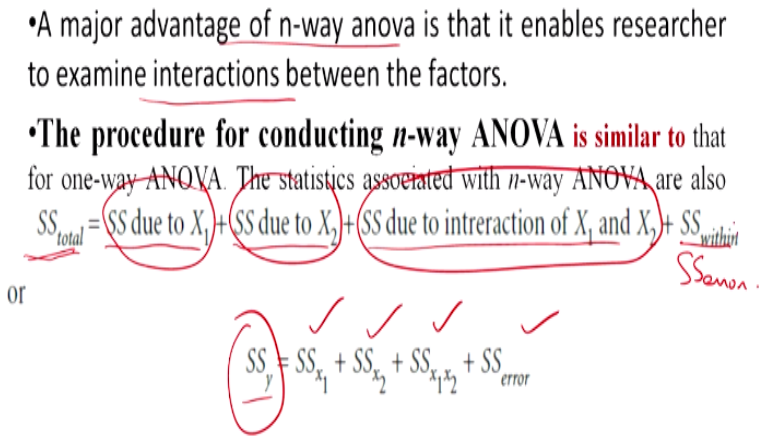
So there are 2 types of you know consumption habits, people eating cereal, no cereal, coffee, and no coffee and these values inside are nothing but a measure of mood quality. So my mood quality is 100, 60, 60, 50, so higher the mood quality, better your health condition is, mental health condition is. So what it is saying is when somebody is consuming cereal and coffee together let us say, the mood quality is 100.

When somebody is consuming cereal but no coffee the mood quality is 60. When somebody is consuming coffee but no cereal it is 60. When both is not, the the mood quality is fifty. Now if I take the SME, the simple main effect, of cereal now we are measuring this SME of cereal with coffee is how much 40, the SME of cereal with coffee. SME of cereal with no coffee without coffee is how much 10, this is 40, this is 10. Now the difference between these 2 is nothing but my interaction effect.

The SME of cereal is larger with coffee than without. Now I want to know whether this difference is a significant difference or not. Similarly for this case. SME of coffee is larger with cereal than without. Now is the effect of coffee is this effect significant effect or not, you want to measure it. How do we conduct this n-way ANOVA? It is very similar to the one-way ANOVA which we had done in the last class.

(Refer Slide Time: 20:15)

Conducting *n*-way ANOVA



The advantage of the n-way ANOVA like I had explained in the factorial design if you remember is that in the factorial design in the statistical design or a truly experimental design, randomness is a very important thing and then when you talk about the interaction effect or the statistical design, we said about one of the factorial designs which helps us to make an interaction effect, which was not possible to check in a Latin square design or a randomized design.

So the advantage is that it helps you to find the interactions. How do you conduct? The procedure for conducting n-way ANOVA is similar to that for the one-way ANOVA. The statistics associated are also the similar, let us see. Now what it is saying? What is SS mean, SS means sum of square. The sum of square total. Now what is this sum of square? Sum of square is a deviation of any value from some other value which could be mean, that is what sum of square means.

The sum of square total = sum of square due to the first independent variable, sum of square of due to the second independent variable + sum of square due to the interaction of x1 and x2 + sum of square within the group or this is the error. So I think you have understood by now. Sum of square within is also a term called a sum of square error. So since you have now 2 factors, earlier you used to have only 1 factor.

So you used to say sum of square due to X that is between + sum of square within, but now since this between has 2 groups, there are 2 groups. So you will say not only due to X1 but

also due to X2 and plus the interaction effect of these 3, all the 3 taken, is very simple very very very simple just understand. SSy is the total = x1, x2, x1x2 + error, so this error, within.

(Refer Slide Time: 22:19)

- A greater mean difference in the levels of X_1 and a larger SSx_1 reflects a larger effect of X_1

- The same is true for the effect of X_2.

- The **larger the interaction** between X_1 and $X2$, **the larger** SSx_1x_2 will be, on the other hand, if X_1 **and** X_2 **are independent, the value of** SSx_1x_2 **will be close to zero.**

A greater mean difference in the level of X1 and a larger SSx1 reflects a larger effect of X1. What does it mean? A greater mean difference in the levels of X1, so that means X1 if there is a value we saw the difference we showed in the last matrix and a larger sum of square reflects that X1 has got a big impact or a significant impact. Similarly the same is true for X2. I have showed you in this, I am talking about this slide. Larger the interaction between X1 and X2, the larger the sum of square x1x2 will be.

On the other hand, if X1 and X2 are independent, there is no relationship, the value of sum of square of x1x2 will be 0 or close to 0, that means the interaction effect does not happen on its own, interaction effects can only happen when there is some relationship or some connection between 2 variables. Had they been completely independent variables, there is no relationship.

For example a living organism and a nonliving organism, if there is no relationship, there will be no interaction, but if there are 2 living organism, maybe one is a fish and other is human, still there can be some interaction, a dog and a human, a dog and a cat, but what is the interaction between let us say a car and a stone, hardly any.

(Refer Slide Time: 23:46)

Steps to follow

- Step 1- The strength of the joint effect of two (or more) factors, or the overall effect.

$$\text{multiple } \eta^2 = (SS_{x_1} + SS_{x_2} + SS_{x_1x_2})/SS_y$$

So steps to follow. Step 1, so how do you conduct this test? The strength of the joint effect of 2 or more factors or the overall effect is to be measured first, how to the multiple eta square. Now what is the formula. Sum of square of x1 + sum of square of x2+ sum of square of x1x2 interaction/total sum of square, this is the formula how you collect and then you check whether it is significant or not.

(Refer Slide Time: 24:12)

Step 2- The significance of the overall effect may be tested by an *F* test, as follows:

$$F = \frac{(SS_{x_1} + SS_{x_2} + SS_{x_1x_2})/df_n}{SS_{error}/df_d}$$

$$= \frac{SS_{x_1,x_2,x_1x_2}/df_n}{SS_{error}/df_d}$$

$$= \frac{MS_{x_1,x_2,x_1x_2}}{MS_{error}}$$

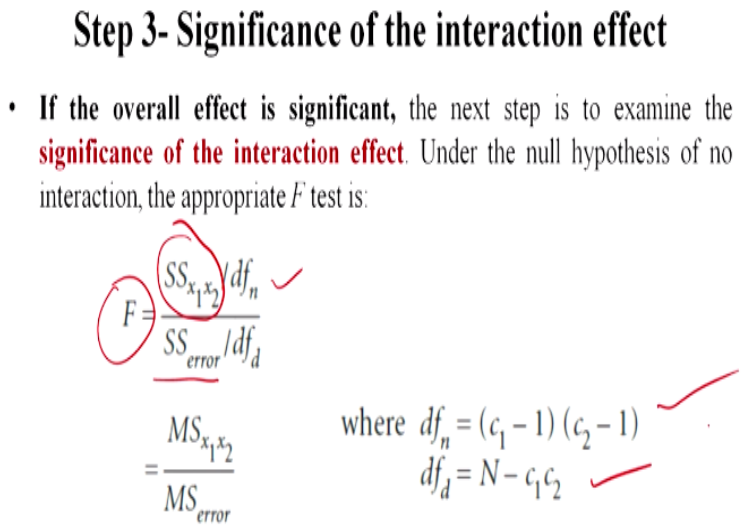
where df_n = degrees of freedom for the numerator
$= (c_1 - 1) + (c_2 - 1) + (c_1 - 1)(c_2 - 1)$
$= c_1c_2 - 1$
df_d = degrees of freedom for the denominator
$= N - c_1c_2$
MS = mean square

How do you check the significance? The significance test is checked by the F value. So to do this what you do is, you take this formula. Sum of square x1+ sum of square x2+ sum of square x1x2 total/degree of freedom in the numerator/sum of square of error divided by the degree of freedom in the denominator. If you remember earlier also the same formula in one-way ANOVA. It was sum of square between you were saying/the degree of freedom between/sum of square within/degree of freedom within.

This is what we were saying, same thing. So by using this formula, you are calculating the mean sum of square of x1x2x1x2/MS error. Now you see the degree of freedom for the numerator is given as how much, now c1c2-1, so there are 2 variables let us say you take x1x2, whatever you c1 you take the row and the column, so the total number of variables into the number of levels x1.

There are let us say 2 brands and 3 cars were there and 4 tiers were there. So the total was 3x4, 12-1, this is what we did. Degree of freedom for the denominator is N, N the total number of cells – c1 x c2, that means number of rows and columns, so in that case how much, we had I have to check, anyway whatever, it is the total number - c1c2. So this is how you go.

(Refer Slide Time: 25:52)

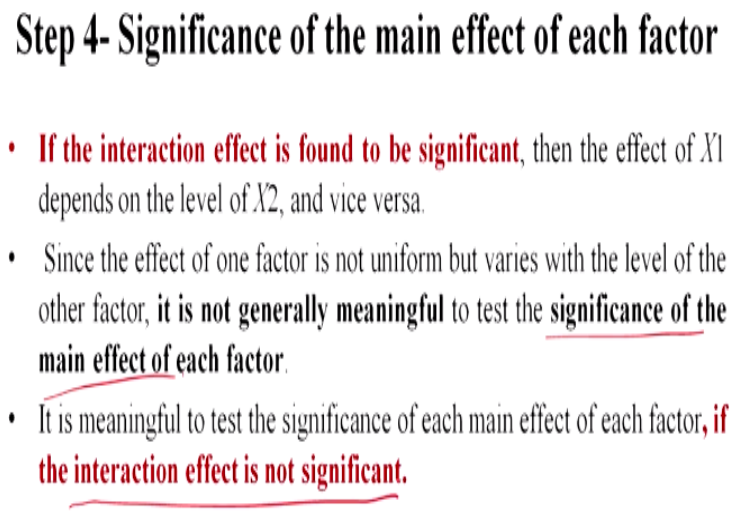


Step three once you have found the significance of the overall model, then you check whether there is any interaction effect or not. Now why this is important? If there is any interaction effect, then understand that main effect do not play any vital role, that means if there is no interaction effect, then only one should go for checking the main effects. Now what is this? If the overall effect is significant, the next step is to examine the significance of the interaction effect.

Under the null hypothesis of the F test is, how do you check the interaction? The F = the interaction sum of square/degree of freedom of the numerator and sum of square within degree of freedom denominator. Now you see what it is saying, degree of freedom numerator

is equal to c1-1, the row minus one, column minus one, same thing, and this is again you do it. Now once this is done, you get an interaction value.

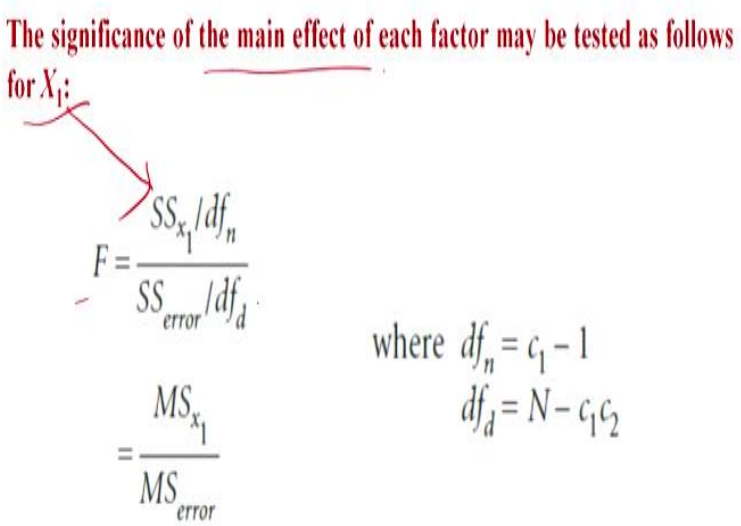
(Refer Slide Time: 26:56)



Now after this, suppose the interaction you have done and then you want to check the main effect. Generally main effect should only be checked when the interaction effect is insignificant. Let us read this. If the interaction effect is found to be significant, then the effect of X1 depends on the level of X2 and vice versa. Since the effect of one factor is not uniform but varies with the level of the other factor, it is generally not meaningful to test the significance of the main effect of each factor.

That means what, if the interaction effect is significant, it makes sense to go for checking the significance of the main effects because it has been seen that the value of X1 depends on the level of X2, that means according to their position, the final the output depends. It is meaningful to test the significance of each main effect of each factor only if the interaction effect is not significant, so this is the thing, this is very important.

(Refer Slide Time: 28:04)



Suppose you found it not significant, then you can check the significance of the main effect of each factor. So there are 2 factors X1 and X2 for example. I am showing you for X1, similarly you do for X2. Now F is equal to what it says, the sum of square of X1, so this is for X1/the degree of freedom/divided by the sum of square error/by the degree of freedom where degree of freedom is at this time it is why it is c-1.

Because only we are talking about one factor, so there is no second factor, so there is only one minus from the same factor. So whatever the c1 number of levels, c1 the number of levels, so there are 3 levels so 3-1, 4 levels 4-1 whatever it is. I hope you are pretty clear with it now what is one-way ANOVA and what is the meaning of analysis of covariance which is ANCOVA and what is n-way ANOVA and how do you calculate by hand and how do you do it through SPSS.

I will show you to in the next lecture. So I hope you enjoyed this lecture and may be in the next lecture, we will learn how to do that and have a little more grasp on the subject. So thank you very much.