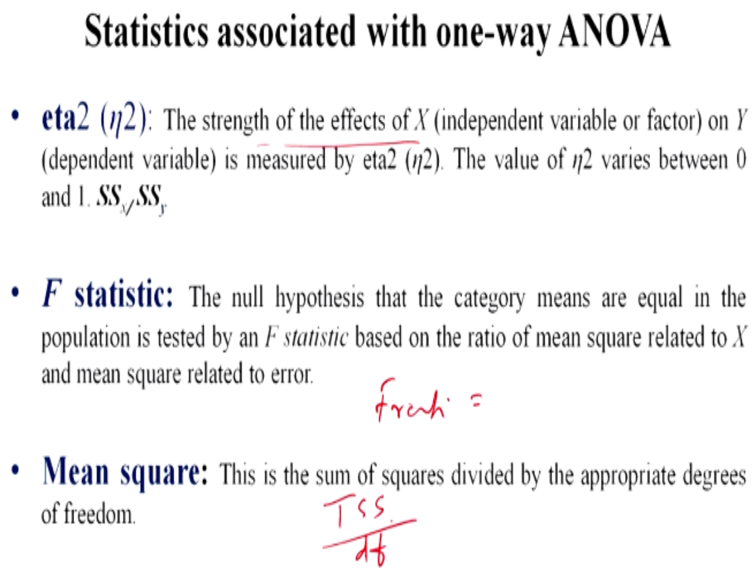
Marketing Research and Analysis-II
(Application Oriented)

Prof. Jogendra Kumar Nayak
Department of Management Studies
Indian Institute of Technology – Roorkee

Lecture – 34
Conducting one-way ANOVA

Welcome friends to the class of Marketing Research and Analysis. We will continue from the last class where we had stopped and in the last class, we had started with a technique called the analysis of variance and analysis of covariance. So we understood the statistics behind it and what is the philosophy of the analysis of variance and covariance. So these are the statistics that are involved.

(Refer Slide Time: 00:55)

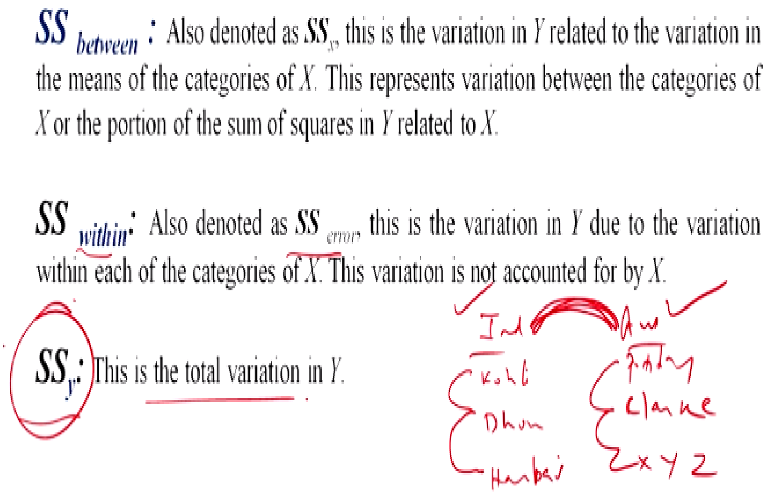


So we said well this is something that the eta square which measure the strength of effect of the independent variable on the dependent variable. Then we said something called F ratio which helps in testing the hypothesis by measuring the significance. Now I am sure you must be aware by this time what significance means, though when we measure significance or test of significance, the researcher rejects or accepts the hypothesis on the basis of this level of significance.

So if you remember when we have a calculated value and we compare it with the critical value or the table value and if it is less or more, accordingly we reject or accept a hypothesis, null hypothesis. Mean square is nothing but the total sum of square, may be divided by the degree of freedom. Now what is degree of freedom, I have explained earlier also. Degree of freedom means, suppose the sample size is 10, then n-1 always is the degree of freedom, that is 9, why it is not 10 and why it is 9, because the tenth one is redundant.

We know that if 9 values are given and the average of the 10 is given to us, then automatically the tenth will be known to us, so that is why it is the degree of freedom test tells us the last one is unnecessary and we do not require.

(Refer Slide Time: 02:30)

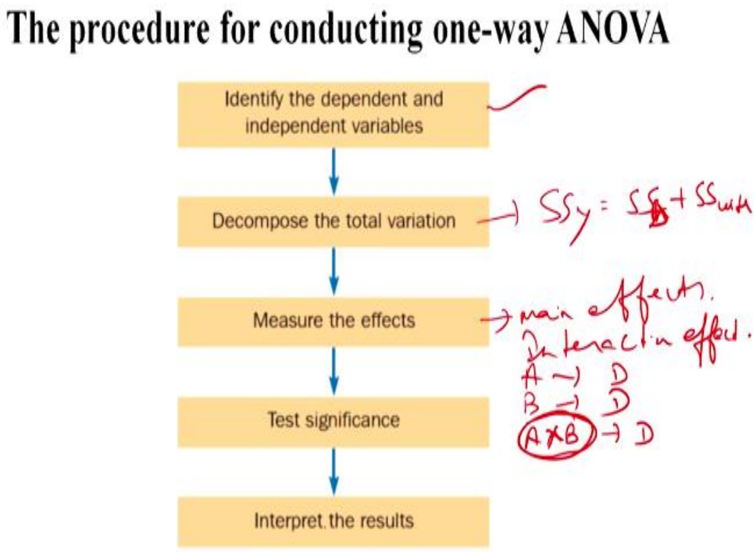


Sum of square between and sum of square within, so what is this we said. Let us say there are, let us say 2 teams India, Australia for example and there are players so Kohli, Dhoni, and Harbhajan and in Australia we have Ponting, we have Micheal Clarke let say and we have somebody let us say a some other batsman, I do not remember anybody’s name, anybody for that let us say XYZ.

Now the difference India and Australia, these are 2 groups, the variation in between these 2 groups which is denoted as the sum of square between is the difference between the team Australia and India’s performance. So this is what that actually we are trying to test. During a hypothesis test we are trying to see whether 2 groups are same or different, so the means of these 2 groups are different or same.

So but if there is a difference between the performance of Kohli and Dhoni and Harbhajan are similarly in between these players, this is not of our interest but this is that is why it is called the error, the sum of square of error. So this sum of square within or sum of square of error is the variation in Y the dependent variable due to the variation within the categories of X, that means this is one category, this is another category and the difference within these categories is what gives us the sum of square of error and total sum of square is given by the total variation is the total sum of square.

(Refer Slide Time: 04:12)



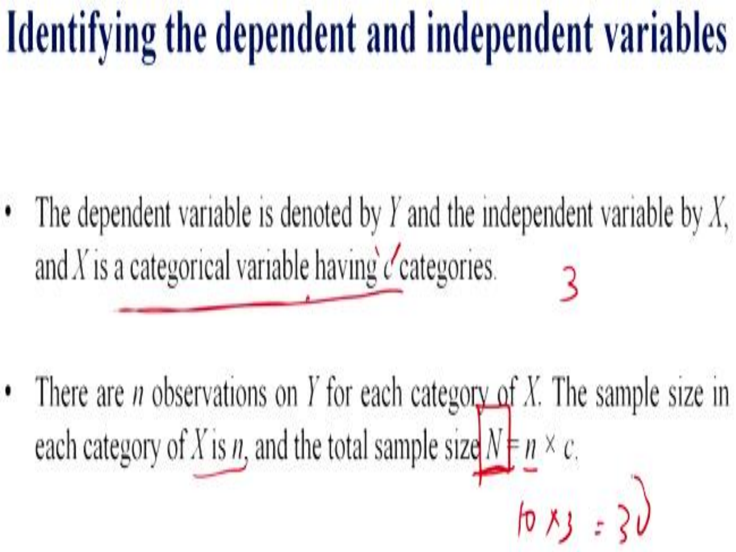
Now how do you proceed with the one-way ANOVA. So first you identify the dependent and the independent variable. So what are you measuring is your dependent variable and what is effecting it is the independent variable. Now find the total variation, the total variation is the sum of square that you just saw here, this one, total variation sum of square Y. Now this is nothing but sum of square of X, that means within the group or you can write sum of square between let us say plus sum of square within.

Now after doing this, you measure the effects. Now what are the effects? Now there are 2 types of basically effects, one is the main effects and the other is called the interaction effect. So you have studied in factorial design, if you go back you will understand, the main effects are the effects of a particular group and the interaction effects are when due to the presence of 2 different categories the kind of change in the score is what we call as the interaction effect.

So it is an interaction between because of the presence of A and B together we have an interaction effect. So let us A effects D is the main effect, B effects D is the main effect, A and B together, A x B, this multiplied value effects D, now this A x B, this impact is my interaction impact. So I have explained this. Now once you do this, you check it, then you measure the significance test and finally you interpret the results.

Now testing the significance means, you try to calculate the F ratio and then you try to say whether the null hypothesis is to be rejected or not rejected. Now these are the steps that you follow.

(Refer Slide Time: 06:08)

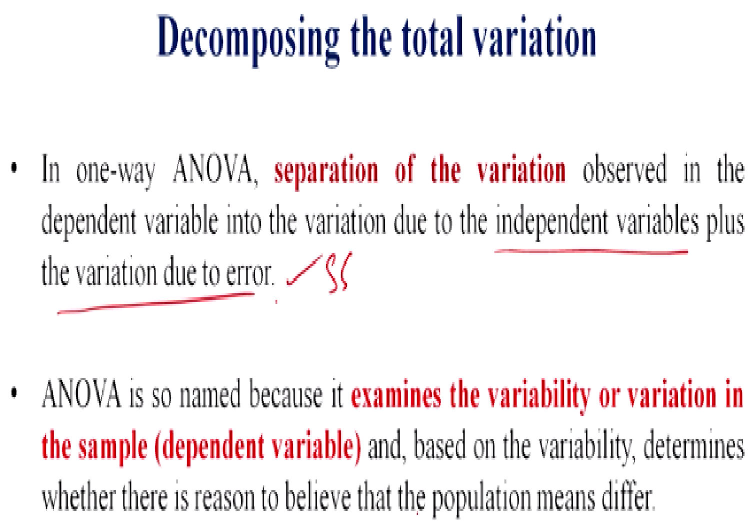


The dependent variable and the independent variable is given as Y and X as it says. The X is a categorical variable having number of categories c categories it is said here. So for example we have as I said let us say we want to see the effect of gender on the academic score. Now gender is my categorical variable independent variable and exam score is my dependent variable. So this has how many categories, 2, male, female.

Income, the effect of the income on let us say the score of the student. Now from what background family he comes from a high income family, low income family, middle income family and his level of the score he is getting in the exam. Now that is let us say 3 categories, let us say you have 3 categories high, medium, low similarly. There are n observations on Y for each category of X, how many observations you are taking for that is what you say.

The sample size in each category is n, that means in particularly 1 group how many respondents you have chosen and the total sample is how much. So capital N is equal to how much, now in each group what is the N into the number of groups, all categories. So let us say there are 10 in one group and there are 3 groups, then how much is the total N, 30.

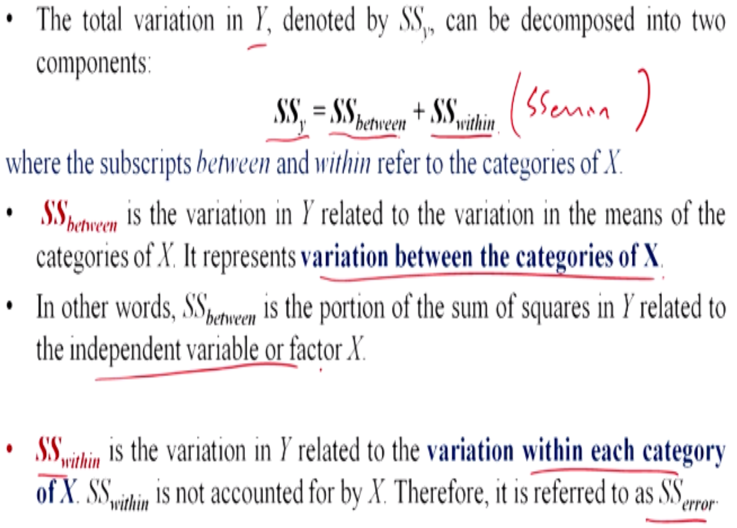
(Refer Slide Time: 07:32)



How do you decompose the total variation? In one-way ANOVA, separation of the variation, you separate the variation observed in the dependent variable into the variation due to the independent variables plus the variation due to error. Now this error is a sum of square error, you are talking about the error which is the within and the sum of square between the groups, so between the independent variables, between the 2 categories let us say.

So ANOVA is so named because that examines the variability or variation in the sample and based on the variability determines whether there is a reason to believe that the population means differ. So it says by understanding or calculating the score and finally the F ratio you can try to infer whether there is difference in the population means, whether the sample groups are coming from the population or they are not coming from the same population, this is that it measures.

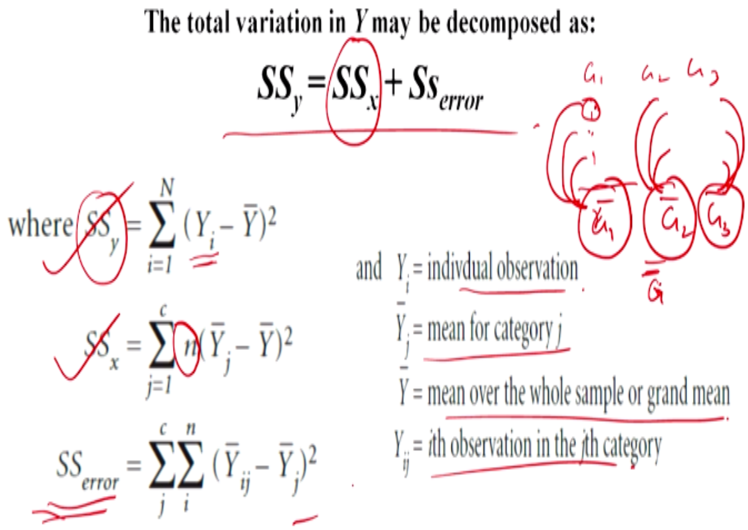
(Refer Slide Time: 08:43)



So total variation Y (denoted by SS_y), this is equal to sum of square between (SS_b) the sum of square within (SS_w) this is what I have said, so this is also called a sum of square error. Sum of square between is the variation in Y related to the variation in the means of the categories of X, it represents variation between the categories of X. In other words, it is the portion of the sum of squares in Y related to the independent variable and similarly therefore this is related to the variation within the category and referred to as error.

So just imagine just understand 2 things, one between the groups, so that is why it is called between sum of square between which is the sum of square between this one and this is the sum of square within which is the within, inside the groups, and that is not of interest and that is the error.

(Refer Slide Time: 09:31)



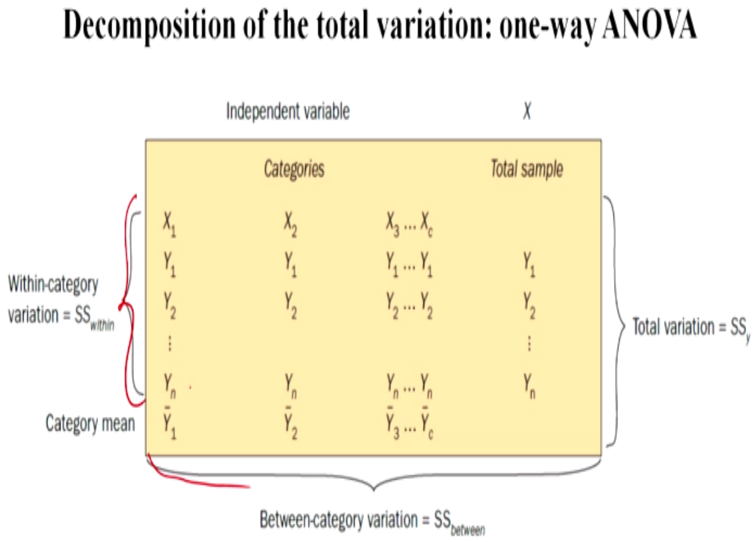
So this is how the total variation comes. $SS_y = SS_x + SS_w$. Now how Y, SS_y you calculate. Now this is the individual observed value minus, you see Y_i is the individual observation and Y bar, this is the Y bar, mean over the whole sample or grand mean. So we will see what is this grand mean and this is the total variation. So let us take a table and you try to find out let us say this is group 1, groups 2, group 3 and you have some values.

So you have a group 1 mean, group 2 mean, group 3 mean and the grand mean your group 1 + group 2 + group 3, mean of these 3 divided by 3, so this gives you the grand mean. So every value minus this, this minus this square plus this minus this square plus this minus square this square plus this minus this square and when you do it completely, so this gives me the sum of square of Y.

Sum of square of X is given by the X means this one the between, thus integration summation of Y bar that means, what is Y bar j mean for the category X, so this one this value, this value, and this value, mean for category j minus the grand mean square multiplied by the number of observations. Why this is important, suppose it would have been same, no issues, but many a times the groups may be unbalanced.

So one group might have 10 players, some group might have 12 players because for some reason. So therefore you need to multiply it with this n. Sum of square within or error is given as the summation of Yij, now what is Yij, the ith observation that means here this one minus the mean for category, so this minus this, this minus this, this minus this, this minus this, similarly this minus this, this minus this, and this minus this and goes on. So when you take these values, this given you the sum of square of error.

(Refer Slide Time: 11:53)



Now this is how the table looks like. So total variation is sum of square of Y, between category this is the between the groups, so these are between the groups, and this is within the categories. So these are within the categories.

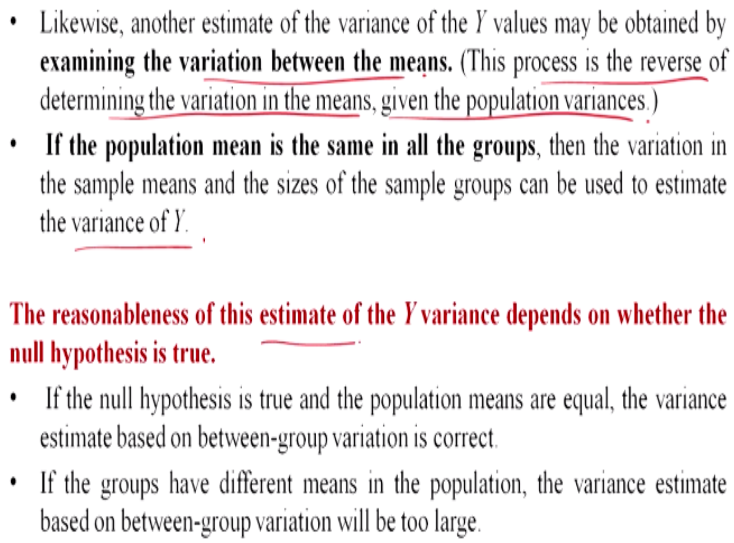
(Refer Slide Time: 12:06)

- In ANOVA, we estimate two measures of variation: within groups (SS_{within}) and between groups ($SS_{between}$).
- **Within-group variation** is a measure of how much the observations, Y values, within a group vary. This is used to estimate the variance within a group in the population.
- It is **assumed that all the groups have the same variation** in the population. But because it is not known that all the groups have the same mean, we cannot calculate the variance of all the observations together.
- The **variance for each of the groups must be calculated individually**, and these are combined into an **'average' or 'overall' variance.**

In ANOVA, we estimate two measures of variation, within groups and between groups. So this is I think we have discussed, so I am just moving to the next one. This is important, it is assumed that all the groups have the same variation, this is called the homogeneity of variance, which is done through a test called the Levene's test in the population, but because it is not known that all the groups have the same mean, we cannot calculate the variance of all the observations together.

Just understand that this tells you that when you want to compare the different groups, you have to assume or the assumption is that the groups have the same variation, so the variance among the groups is the same, that means it is not different. So the variance for each of the groups must be calculated individually and these are combined into an overall variance.

(Refer Slide Time: 13:07)



Likewise another estimate of the variance of the Y values must be obtained examining the variation between the means. This process is the reverse of the determining the variation in the means, so it gives the population variances. Now what this means? If the population mean is the same in all the groups, then the variation in the sample means and the size of the sample groups can be used to estimate the variance of Y, Y is my dependent variable.

It says the reasonableness of the estimate of the Y depends on whether the null hypothesis is true or not. Now if the null hypothesis is true, the population means are equal and the variance estimated based on the between-group variation is correct, what it means? If my null hypothesis is true and the population means are equal, then variance estimate based on the between the groups that means group 1, group 2, group 3, the estimate between the groups variation is correct.

But if the groups have different means, that means mu1 is not equal to mu2 is not equal to mu3, the variance estimate will be between the group variation will be too large. So it actually explains you what is the effect of the variance on the overall study.

(Refer Slide Time: 14:32)

Measuring the effects

- The effects of X on Y are measured by SS_x.
- Because SS_x is related to the variation in the means of the categories of X, the relative magnitude of SS_x increases as the differences among the means of Y in the categories of X increase.
- The relative magnitude of SS_x also increases as the variations in Y within the categories of X decrease.

So how do you measure? The effects of X on Y that means the independent variable on the dependent variable, are measured by this factor called the sum of square of the variance within between the categories. Since it is related to variation in the means of the categories of X, the relative magnitude of SSx increases as the difference among the means of Y in the categories of X increase.

So well let us understand this. As SSx increases, that means the between the group variance increases, the differences among the means of Y in the categories of X also increase. The relative magnitude SSx also increases as the variations in Y within the categories of X decrease. So 2 things, when the variance between the group decreases, it is good. When the variance within the category or the error variance decreases, it is also good.

(Refer Slide Time: 16:01)

The strength of the effects of X on Y is measured as follows:

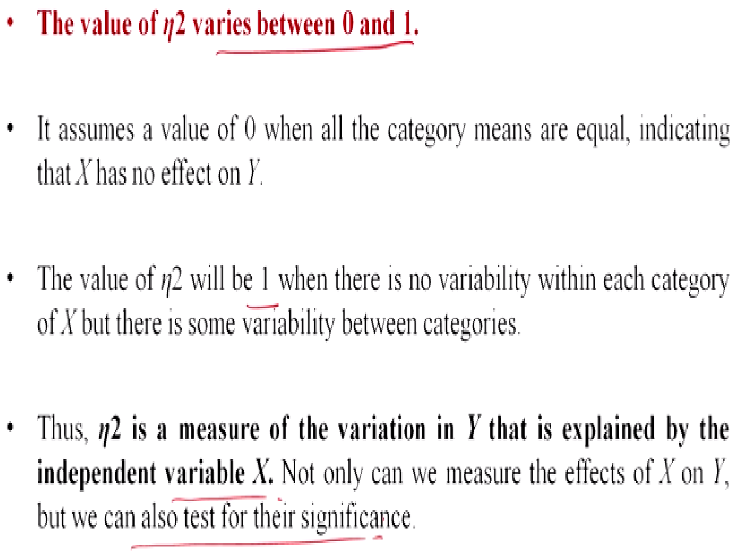
$$\eta^2 = \frac{SS_x}{SS_y} = \frac{SS_y - SS_{error}}{SS_y}$$

$$\frac{\text{Variance between } (SS_x)}{\text{Total variation } (SS_y)}$$

So how do you measure the strength. Now this eta square is given as sum of square of X that is the variance between the groups divided by the variance within by the total variance, so this you can understand variance between which is SSx/total variance, the one which you just discussed here, here we were discussing somewhere the estimate of Y variance, so by total variance which is SSy.

So if you divide SSy that means you can write it as total error - within error = between error, so this same thing, so instead of writing SSx, I am writing it as SSy-SS error/the total error.

(Refer Slide Time: 16:36)



The value of eta square lies between 0 and 1. It assumes a value of 0 when all the category means are same or equal, that means the null hypothesis is correct, is to be accepted indicating that X has no effect on Y. The value of eta square will be 1 when there is no variability within each category, that mean within the error term is completely 0, but there is a variability between the categories and this is what we want during ANOVA.

We want that the between group variation should be maximum and the within group variation should be minimum, almost 0, not 0, but almost 0 we can say. Thus the eta square is a measure of the variation in Y that is explained by the independent variable X, not only can we measure the effects of X on Y, we could also test for the significance through the F ratio which we have already discussed.

(Refer Slide Time: 17:39)

Testing the significance

Null hypothesis that the category means are equal in the population.

$$H_0: \mu_1 = \mu_2 = \mu_3 = \ldots = \mu_c$$

- Under the null hypothesis, SS_x and SS_{error} come from the same source of variation.
- In such a case, the estimate of the population variance of Y can be based on either between-category variation or within-category variation.

So null hypothesis that the category means are equal is given like this. So under this null hypothesis SSx between and the within come from the same source of variation from the different groups and the members within the groups. In such a case, the estimate of the population variance of Y can be based either between category variation or within category. So this has to be more, this has to less, so less more.

(Refer Slide Time: 18:13)

The estimate of the population variance of *Y*

$$S_y^2 = \frac{SS_x}{c-1}$$

= mean square due to X

= MS_x

or

$$S_y^2 = \frac{SS_{error}}{N-c}$$

= mean square due to error

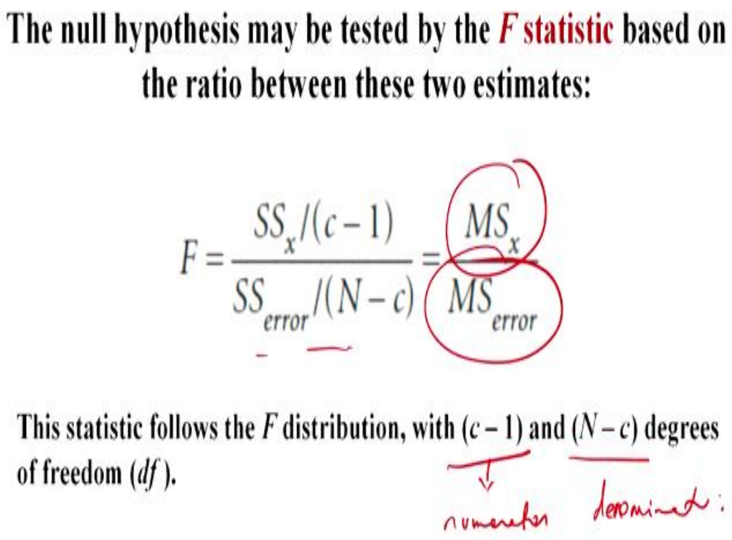
= MS_{error}

$$F = \frac{MS_x}{MS_e}$$

The estimate of the population variance of Y is given as like this. So Sy square variation is given by SSx, sum of square between, divided by the c-1, now c-1 actually is nothing but my degree of freedom. Now what is the degree of freedom? The number categories or the number of columns minus 1, so this gives me the mean sum of square due to X or this is called as MSx.

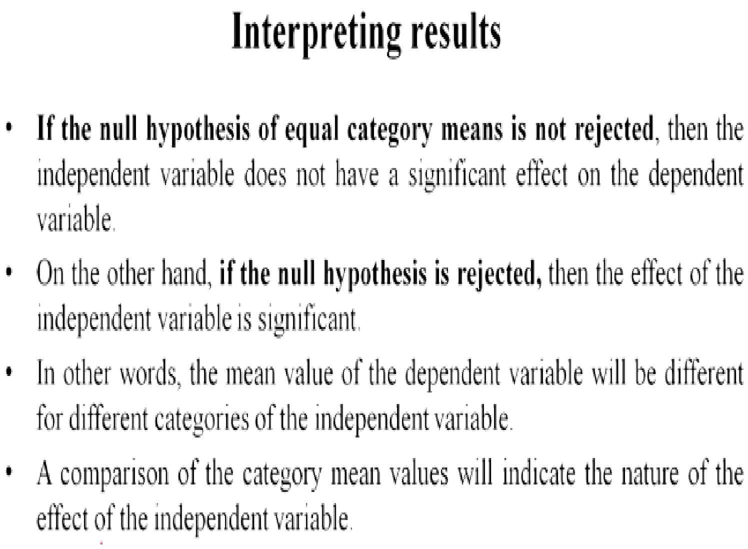
Similarly we want to calculate the sum of square that means the variation the mean sum of square within the groups, so mean sum of square within, so this is equal to sum of square of the within or error divided by what N-c, now what is N, N is the number of number of respondents in each category x c = N we had seen earlier. Now this minus the c, that means what for example whatever the total N, the total size - the number of columns, so this is called my MS error. Now this ratio of this MSx/MSe is what is my F ratio.

(Refer Slide Time: 19:31)



So F = SSx/c-1 which is MSx/MSe which is my MS error. This statistics follows the F distribution with c-1 and N-c degrees. So if you look at the F distribution table, this is my numerator and this is my denominator of degrees of freedom.

(Refer Slide Time: 19:59)

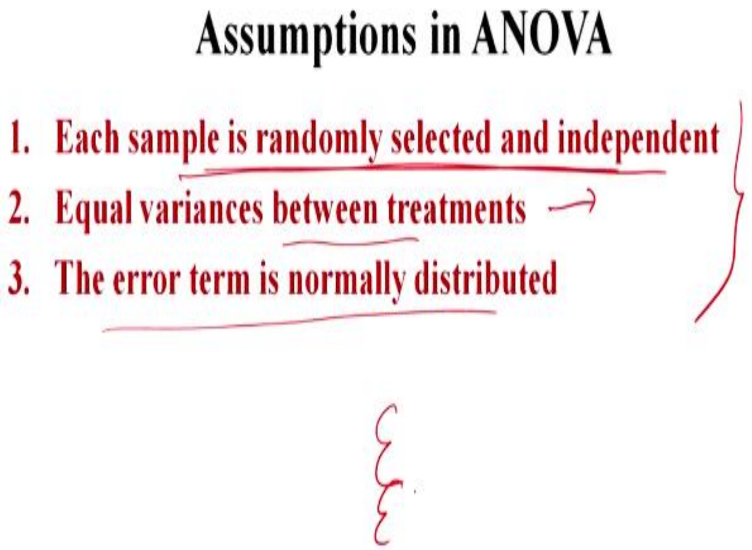


Now how do you interpret the results? If the null hypothesis of equal category means is not rejected, then the independent variable does not have a significant effect, is not rejected means we will say that they are all equal. On the other hand if the null hypothesis is rejected, that means the null hypothesis is what that the means are all equal, if that is rejected, then the effect of the independent variable is significant.

That means we will say that the independent variable in this case for example what we were discussing earlier is that gender has a importance or plays a role on the score that students own in their class exams or the kind of family they come from, income category plays a role in the kind of results that they get during the exams. So the mean value of the dependent variable will be different for different categories of the independent variable.

A comparison of the category mean values will indicate the nature of the effect of the independent variable. So when we compare the mean values, it can tell us what effect it has got finally on the dependent variable.

(Refer Slide Time: 21:08)



So these are some of the assumptions. Each sample is randomly selected, so because it comes from that experimental design and in that experimental design, it is a statistical design but it is purely experimental design, so randomness has to be one of the most important criteria. So there is a randomness here. Equal variances, we just talked about the homogeneity of variance.

So between the groups or the treatment levels, there has to be an equal variance, though if the variance is different, then we cannot compare it. The error term is normally distributed, so within the group variation is normally distributed.

(Refer Slide Time: 21:51)

EXAMPLE (ONE WAY ANOVA)

As head of a department of a consumers' research organization, you have the responsibility for testing and comparing lifetimes of four brands of electric bulbs. Suppose you test the lifetime of three electric bulbs of each of the four brands. The data is shown below, each entry representing the lifetime of an electric bulb, measured in hundreds of hours:

Brands

A	B	C	D
20	25	24	23
19	23	20	20
21	21	22	20

* Can we infer that the mean lifetime of the four brands of electric bulbs are equal?

Now let us take an example. This is a case, there are 4 brands. See the head of a department of consumers' research organization is testing and comparing lifetimes of 4 brands of bulbs. Suppose you test the lifetime of 3 electric bulbs of each of the 4 brands, so 3, 3, bulbs of each he has taken. A 3 bulbs, B, C, D. The data is shown below each entry representing the lifetime of an electric bulb measured in 100 of hours, so 2000, 1900, 2100, it goes on. Can you infer that the mean lifetime of the 4 brands of electric bulbs are equal?

Now if would have done a t test, how you would have done? You would have done the comparison between A and B, A and C, A and D, B and C, B and D, C and D. So you see there are numbers of combinations now, you have 6 possible combinations, so that would have become a cumbersome factor. Now we will do it through the one-way ANOVA.

(Refer Slide Time: 22:50)

SOLUTION

The null hypothesis is that the mean lifetime of the four brands of electric bulbs are equal, i.e.,

$$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4.$$

Let $\bar{X}_1$, $\bar{X}_2$, $\bar{X}_3$ and $\bar{X}_4$ denote the mean lifetime of Brand A, B, C and D respectively and be the overall grand mean.

Then,

X_1	X_2	X_3	X_4
20	25	24	23
19	23	20	20
21	21	22	20
$\bar{X}_1 = 20$	$\bar{X}_2 = 23$	$\bar{X}_3 = 22$	$\bar{X}_4 = 22$

What is the null hypothesis? The mean lifetime of the 4 brands of the electric bulbs are equal. So what is my null hypothesis the mu1 = mu2 = mu3 = mu4 correct. Now what it says is let X1, X2, X3, X4 denote the mean lifetime of brand A, B, C, and D. So these are the brands right, so the mean lifetime so 20, 19, 21. So what is the average, so let say this is the mean for the first group is the first one is 20, the mean for the second one is 23, third one 22, fourth one 22.

(Refer Slide Time: 23:32)

And,

$$\text{Grand mean, } (\bar{\bar{X}}) = (\bar{X}_1 + \bar{X}_2 + \bar{X}_3 + \bar{X}_4)/4$$

$$= (20 + 23 + 22 + 21)/4$$

$$= 21.5$$

So what is the grand mean X1+X2+X3+X4/4 so that comes 21.5. So now we will be doing 2 things. One we will be comparing individually each value to the grand mean. When this value minus the grand mean square plus, similarly from this, this, this every individual value and we also have to find out for the total variation, for individually the between the groups you will have to find the variation that between this mean minus the grand mean square plus this mean minus the grand mean, this from the grand mean, this from the grand mean.

Within the error or the sum of square of error which is within the group error that you will find out from this minus this, this minus this, this minus this, similarly this minus this, this minus this, this minus this, so there are 3 things that you need.

(Refer Slide Time: 24:23)

Calculating Sum of Square total (SST)

X	$\bar{X}$	$(X-\bar{X})$	$(X-\bar{X})^2$
20	21.5	-1.5	2.25
19	21.5	-2.5	6.25
21	21.5	-0.5	0.25
25	21.5	3.5	12.25
23	21.5	1.5	2.25
21	21.5	-0.5	0.25
24	21.5	2.5	6.25
20	21.5	-1.5	2.25
22	21.5	0.5	0.25
23	21.5	1.5	2.25
20	21.5	-1.5	2.25
20	21.5	-1.5	2.25
	SST		$\sum(X-\bar{X})^2 = 39$

IIT ROORKEE NPTEL ONLINE CERTIFICATION COURSE

Now let us calculate. So what is it saying? So these are the life 20, 19, 21 hours, number of hours, X double bar is my grand mean is this much. So first let us calculate from individually each point from the each group. So the first value 20-21.5, this much, nineteen this, so individually we have subtracted from the grand mean. So we get x-x bar, and x-x bar square we get which gives total variation, so this is coming x-x bar square is equal to 39.

(Refer Slide Time: 25:00)

Calculating Sum of Square Within

A		B		C		D	
X_1	$(X_1-\bar{X}_1)^2$	X_2	$(X_2-\bar{X}_2)^2$	X_3	$(X_3-\bar{X}_3)^2$	X_4	$(X_4-\bar{X}_4)^2$
20	0	25	4	24	4	23	4
19	1	23	0	20	4	20	1
21	1	21	4	22	0	20	1
$\bar{X}_1=20$	$\sum(X-\bar{X}_1)^2 = 2$	$\bar{X}_2=23$	$\sum(X-\bar{X}_2)^2 = 8$	$\bar{X}_3=22$	$\sum(X-\bar{X}_3)^2 = 8$	$\bar{X}_4=21$	$\sum(X-\bar{X}_4)^2 = 6$

Sum of Square Within (SSW) = 2+8+8+6 = 24

Now let us calculate sum of square within. Now how do you calculate sum of square within? Now you have X1, X2, X3, X4. Now X1 have 20, 19, 21 and X1 from the grand mean square if you would subtract, so 20, what is the grand mean, within so you have to calculate from the group mean, sorry not grand mean group mean. So 20-20 square, 19-20, 21-20 similarly, what is the group mean for this 23. So 25-23 square, 23-23 square, 21- so goes on for all.

So you have calculated the total. So sum of square within is equal to how much 2+8+8+6 so that is equal to 24. Now you have got a total sum of square, you have got the within the sum of square, you can just subtract total sum of square minus within sum of square and you can get the between sum of square. Else, you can also calculate, it is also given.

(Refer Slide Time: 26:06)

We know that.

Sum of square total = Sum of square within + Sum of square between

- SST = SSW +SSB
- SSB = SST – SSW

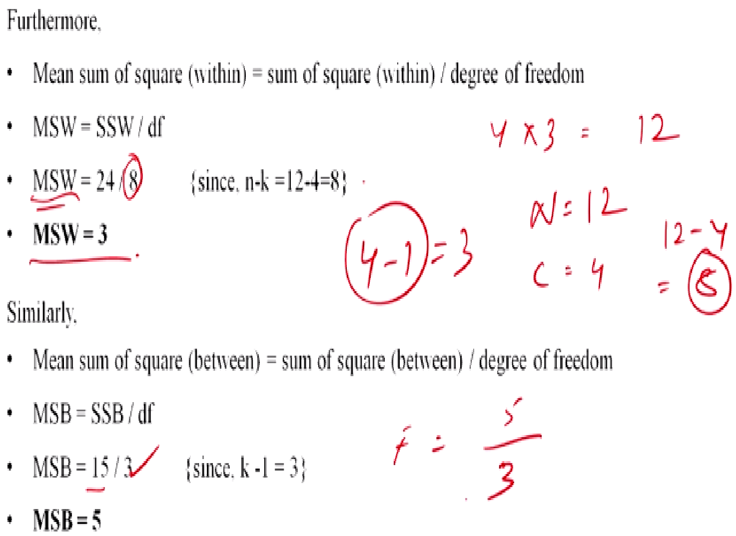
  = 39 – 24

  = 15

So now SST = SSW + so gives you 39-24, so this is 24, this was 39 and you can calculate it, is equal to 15. So now if you directly want to calculate, you can do that. Let us see how you can do, if you want you can calculate. For example this is the group mean. So you can say 20, now if I want to let us write it down, 20, 23, 22, 21 and you have the grand mean, this is the group mean.

So this is the grand mean how much 21.5, 21.5, 21.5, 21.5. Now if you subtract how much is coming 20-21.5 = 1.5 square x how many 3, 3 values are there. Similarly here also 3 x, twenty one point five to twenty two, 1.5 square right. This is 3 x 0.5 square. If you take this total value and you add it up you should get fifteen, so fifteen fifteens are 2.25, so 22.5 x3 that gives you 6.75, this also 6.75, this is 0.25 x 3 = 0.75, so this is again 0.75, so 6 6 12 13.5, so it is 15 same thing right. So we have got either you subtract or you calculate same thing.

(Refer Slide Time: 27:50)



Now mean sum of square is sum of square by the degree of freedom. So let us calculate for each. So mean sum of square within the groups is equal to how much now 24 the total divided N, now how much is the N here, now let us see within this you have 4 groups, each group had 3 values, so total was 12. Now N = 12 and c the number of columns or the categories is equal to 4, so this 12-4 is what is my degree of freedom, so is equal to 8, so that gives me MSW = 3.

Similarly mean of sum of square between = sum of square between/degree of freedom, so that is equal to how much 15 divided by the columns, so how many, you have 4 columns, minus 1, so that is equal to 3. So this is equal to 5.

(Refer Slide Time: 28:59)

Therefore,

- F ratio = MSB /MSW
  = 5/ 3
  = **1.67**
- From the F- table, the value of F for (3, 8) d.f. and at 5% level of significance is 4.07. Since the computed value of F= 1.67 is less than the table value of F= 4.07, therefore , we accept our null hypothesis.
- Hence, the difference is insignificant and we can infer that the average lifetime of different brands of bulbs are equal.

So now F ratio F = mean sum of square between (MSB) that is 5/ mean sum of square within (MSW) 3, so 1.67. From the F table, the value for F for 3 and 8 degrees of freedom and 5% level of significance is how much, 4.07, the computed value is 1.67. So that means what? You cannot reject the null hypothesis, rather you accept the null hypothesis and you say that means the difference is insignificant.

That means there is no difference between the means of all the 4 brands of the electric bulbs and therefore we will say that the life of the bulbs are all equal. So this is what we have discussed today. In the next lecture, we will talk about the two-way ANOVA, then we add more than one ANOVA and we will try to solve some problems on that also. I hope today you are clear how you should proceed with one-way ANOVA.

It not only helps you to understand the main effects but it also helps you to understand the interaction effects between the variables. So how this thing is useful and how you can use it while writing a research paper or your thesis is very vital, and I hope it is clear, and we will continue further in the next lecture. Thank you so much.