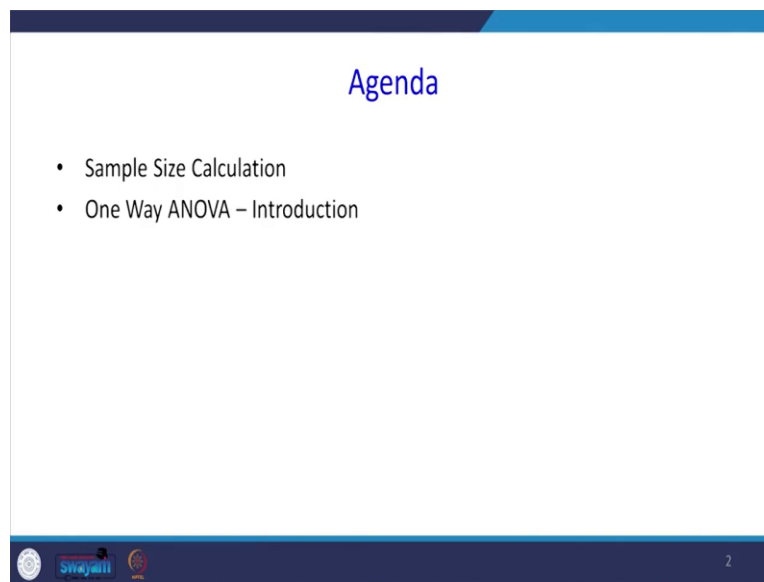


Data Analytics with Python
Prof. Ramesh Anbanandam
Department of Management Studies
Indian Institute of Technology – Roorkee

Lecture – 23
ANOVA- I

Welcome students today we are, we will continue with the sample size calculation. After that we are going to see very important topic that is analysis of variance.

(Refer Slide Time: 00:37)



Today's lecture planet sample size Calculation and one way Anova.

(Refer Slide Time: 00:41)

Determining Sample Size when Estimating μ

- Z formula $Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$
- Error of Estimation (tolerable error) $E = \bar{X} - \mu$
- Estimated Sample Size $n = \frac{Z^2 \sigma^2}{E^2} = \left(\frac{Z \sigma}{E} \right)^2$
- Estimated σ $\sigma \approx \frac{1}{4} \text{range}$

$Z \frac{\sigma}{\sqrt{n}} = \bar{X} - \mu$
 $\frac{Z \sigma}{(\bar{X} - \mu)} = \sqrt{n}$
 $\frac{Z \sigma}{(\bar{X} - \mu)^2} = n$

Determining sample size with estimating Mu: We know that the Z formula is $(\bar{x} - \mu)$ divided by (σ/\sqrt{n}) . In this Sigma by root n, from this relationship this n is nothing but your sample size when you re-adjust this Mu, re-adjust this from this equations like $(Z \cdot \sigma)/\sqrt{n}$, $(\bar{X} - \mu)$, right then \sqrt{n} , can say $Z \cdot \sigma$ divided by $(\bar{x} - \mu)$, that will be your root n. You square both sides.

So, $(Z \cdot \sigma)^2$ divided by $(\bar{x} - \mu)^2$, that will be your n. That is nothing but this one. So, the numerator this $\bar{X} - \mu$ here we are going to call it as Error of estimation otherwise tolerable error. So, sample size is mu such that Z square, Sigma square, this is sigma square, Sigma square divided by error square ok. Since there, everywhere square is there we can bring it to common.

Many times, the value of population standard deviation may not be known. So there is approximation, one fourth of the range of the data which were collected can be taken as standard deviation.

(Refer Slide Time: 02:16)

Example: Sample Size when Estimating μ

$$E = 1, \sigma = 4$$

$$90\% \text{ confidence} \Rightarrow Z = 1.645$$

$$\begin{aligned} n &= \frac{Z^2 \sigma^2}{E^2} \\ &= \frac{(1.645)^2 (4)^2}{1^2} \\ &= 43.30 \text{ or } 44 \end{aligned}$$

We will do an example, calculating the sample size. So, the permissible error is given. 1. The population standard deviation is 4. You want to conduct 90% confidence level. If it is 90% confidence level, this Z value is 1.645. Then you substitute here. You will get 43.30 that is nothing equivalent to 44.

(Refer Slide Time: 02:42)

Example

$$E = 2, \text{ range} = 25$$

$$95\% \text{ confidence} \Rightarrow Z = 1.96$$

$$\text{estimated } \sigma: \frac{1}{4} \text{ range} = \left(\frac{1}{4}\right)(25) = 6.25$$

$$\begin{aligned} n &= \frac{Z^2 \sigma^2}{E^2} \\ &= \frac{(1.96)^2 (6.25)^2}{2^2} \\ &= 37.52 \text{ or } 38 \end{aligned}$$

We will do another problem to find out the sample size. So, permissible error is 2. Range is given 25. 95% confidence level Z is equal to 1.96, this value which you have to get from the table. So, estimated Sigma is one fourth of the range. So it is 6.25 then substitute Z Square, Sigma square divided by E square we are getting 38.

(Refer Slide Time: 03:08)

Determining Sample Size when Estimating P

- Z formula

$$Z = \frac{\hat{p} - P}{\sqrt{\frac{P \cdot Q}{n}}}$$

- Error of Estimation (tolerable error)

$$E = \hat{p} - P$$

- Estimated Sample Size

$$n = \frac{Z^2 PQ}{E^2}$$

$$\begin{aligned} Z \sqrt{\frac{PQ}{n}} &= \hat{p} - P \\ Z^2 \frac{PQ}{n} &= (\hat{p} - P)^2 \\ \frac{Z^2 PQ}{(\hat{p} - P)^2} &= E^2 \end{aligned}$$

Now, we will see how to find the sample size when estimating the population proportion. If you are estimating the population proportion, the formula for Z is different. \hat{p} - Capital P root of capital P and capital Q by n. This P hat is the sample proportion capital P is the population proportion so the error is P - capital P. The same way what we have done previously. If you for example, when you bring Z root of PQ divided by n is equal to P hat minus P. Square both sides.

Z square PQ by n equal to P hat minus P, whole square. So, this will become Z Square PQ divided by P hat minus P whole square. That is nothing but your n. So, P hat - P we call it as E Square. So, n equal to Z Square PQ divided by E square.

(Refer Slide Time: 04:24)

Example

$$E = 0.03$$

$$98\% \text{ Confidence} \Rightarrow Z = 2.33$$

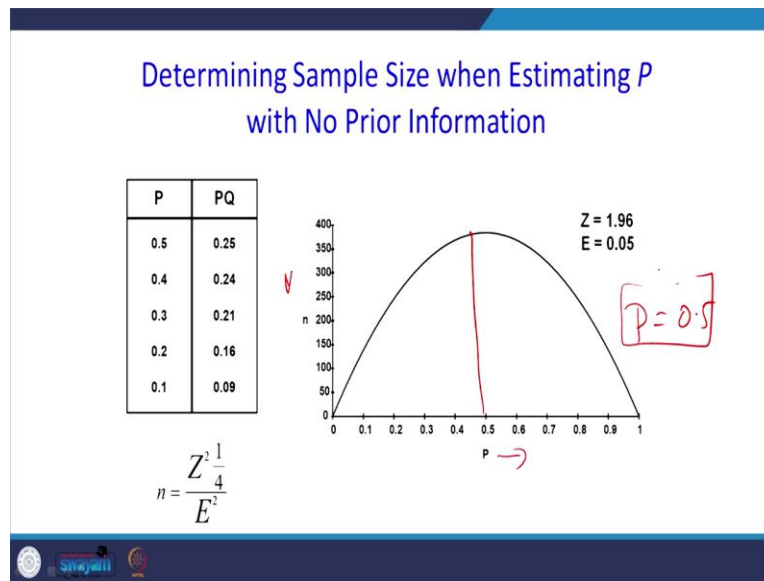
$$\text{estimated } P = 0.40$$

$$Q = 1 - P = 0.60$$

$$\begin{aligned} n &= \frac{Z^2 PQ}{E^2} \\ &= \frac{(2.33)^2 (0.40)(0.60)}{(0.03)^2} \\ &= 1,447.7 \text{ or } 1,448 \end{aligned}$$

We will do one problem here. So the permissible error is given. For 98% confidence level the Z value is 2.33 which you have to get it from the table. Estimated P is given population P is given 0.40. So Q is 0.60. You substitute this value it will be 1448 you see that whenever you go for population testing population proportion generally you ask yes or no type. That time obviously you have to have go for more number of samples. Sometimes what will happen the value of P which were given as 0.4 may not be known to you.

(Refer Slide Time: 05:08)



That time, how to decide the sample size, determining the sample size when estimating P with no prior information. Just look at this. The P is in the x-axis sample size in the y-axis suppose, the value of P equal to 0.5, the maximum sample size is required. So what the logic what we have to understand from here is if you are not knowing the population proportion we have to assume P equal to 0.5. Even the value of P is goes above 0.5 see that the, the n values decreasing.

The value of P is goes below 0.5 that time also the value of p is decreasing. It is better to assume P equal to 0.5 so that you will get maximum number of sample size.

(Refer Slide Time: 06:02)

Example

$$E = 0.05$$

$$90\% \text{ Confidence} \Rightarrow Z = 1.645$$

with no prior estimate of P , use $P = 0.50$

$$Q = 1 - P = 0.50$$

$$\begin{aligned} n &= \frac{Z^2 PQ}{E^2} \\ &= \frac{(1.645)^2 (0.50)(0.50)}{(0.05)^2} \\ &= 270.6 \text{ or } 271 \quad \checkmark \end{aligned}$$

In this situation, Error is given 90% is confidence level Z equal to 1.645 with no prior estimate of P we have to use $P = 0.50$. If we substitute P as 0.50 it will go for 271. The population proportion is not known to you or to take P value equal to 0.5.

(Refer Slide Time: 06:30)

Why ANOVA?

- We could compare the means, one by one using t -tests for difference of means.
- Problem: each test contains type I error
- The total type I error is $1 - (1 - \alpha)^k$ where k is the number of means.
- For example, if there are 5 means and you use $\alpha = .05$, you must make 10 two by two comparisons.
- Thus, the type I error is $1 - (.95)^{10}$, which is .4012.
- That is, 40% of the time you will reject the null hypothesis of equal means in favor of the alternative!

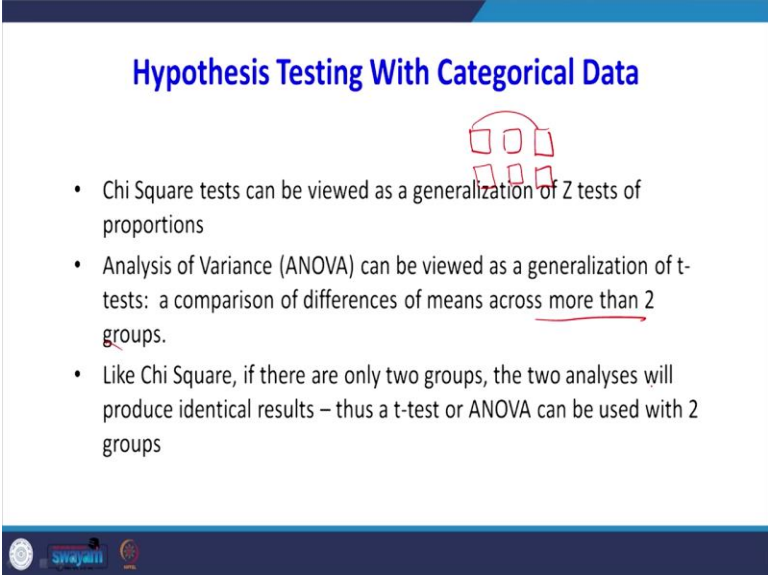


Next will go to next topic is Anova. Why Anova is required? So far we have seen two samples Z Test, two sample T test. Whenever there is a requirement for comparing more than two population. So far, what you have done? First, we have done one population next we have compared two population. If you want to compare more than two population, we can go for Anova. There is a possibility. You can compare 1 and 2 Suppose this is one, this is 2, this is 3, we can compare 1 and 2, 2 and 3, 1 and 3.

1 and 3 there are 3 comparison, but we will not go for that one because there is a reason for that. I will tell you we could compare the means one by one using T test of difference of means because what will happen each test contain Type 1 error. So the total type of error $1 - 1 - \text{Alpha}$ power k where K is number of means. For example, if there are 5 means if you use Alpha equal to 0.05 because $5C_2$. You must, you must make 10 2 by 2 comparison because every comparison, your confidence level is 0.95.

If you are making 10 comparison the overall confidence level is 0.95 power 10. So 1 - confidence level is nothing but error. So, 0.95 power 10 that you substitute for 10 that is nothing but your error. That is 40 % that is, 40 percentage of the time you will reject the null hypothesis of equal means in favour of alternative. That is why we should not go for two samples T test whenever there if you want to compare more than two populations. We should go for Anova.

(Refer Slide Time: 08:32)



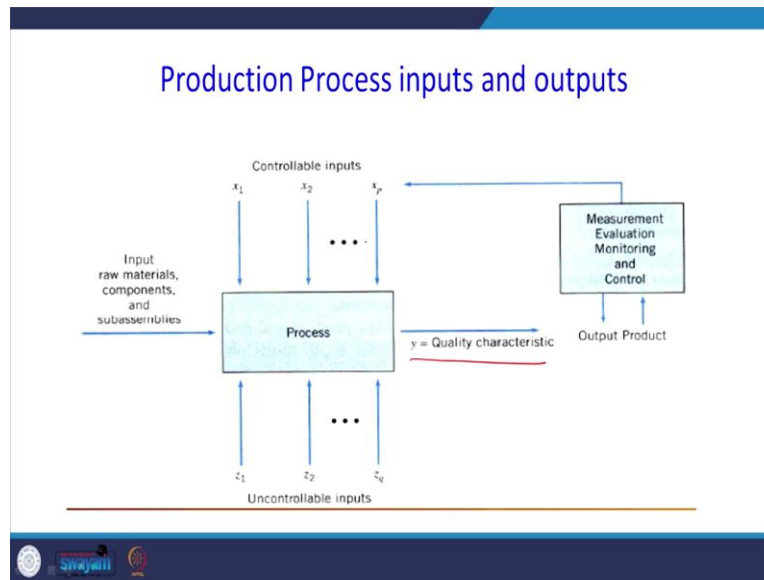
Hypothesis Testing With Categorical Data

- Chi Square tests can be viewed as a generalization of Z tests of proportions
- Analysis of Variance (ANOVA) can be viewed as a generalization of t-tests: a comparison of differences of means across more than 2 groups.
- Like Chi Square, if there are only two groups, the two analyses will produce identical results – thus a t-test or ANOVA can be used with 2 groups

Hypothesis testing with categorical data because we are going to say Group 1, group2, group3. You see we have seen 2 samples Z proportion test. If you want to go for Z proportion test, if you want to go for comparing three population proportion, if you want to compare the proportion of more than two population, we should go for chi square test. Similarly, if you want to compare more than three population mean we should go for Anova.

So, chi-square test can be viewed as the generalization of Z test of proportion. The same way the anova can be viewed as the generalization of T test. The comparison of difference of mean across more than two groups like Chi square if there are only two groups, 2 analysis will produce identical result does a t test and Anova can be used with two groups. There are 2 groups we can go for Anova and a T test. Both will give you the same answer.

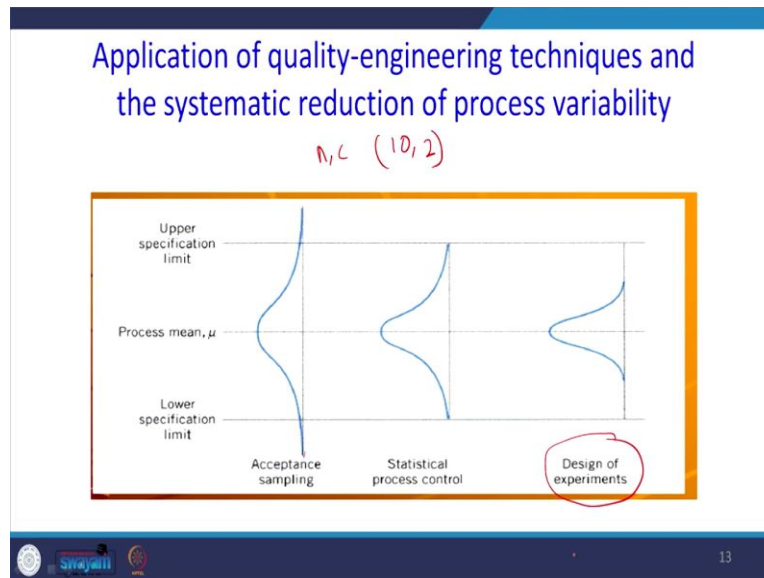
(Refer Slide Time: 09:38)



Why this concept of Anova is more important. Suppose, in the production process, there are different input variables. X_1 , X_2 and X_P . These are controllable inputs. There are uncontrollable inputs, that is, Z_1 , Z_2 , Z_3 . So, the input maybe raw materials, component and subassemblies, output is the quality characteristics. What will happen this quality characteristics, Y is generally affected by this X_1 X_2 X_3 right. See that is output product.

If it is affecting Y we need to find out what combination of X_1 and X_2 will provide better Y value measurement evaluation monitoring and controlling for that purpose Anova is required. So the purpose of Anova is there are many input variables how this input variable is affecting the quality characteristics. For that we can find out with the help of Anova.

(Refer Slide Time: 10:42)



See, when we want to improve the quality Application of quality engineering techniques and systematic reduction of process variability, because we want to improve the quality, is nothing but reduction of the variance. To go for acceptance sampling there will be lot of variance, because the acceptance sampling is a suppose, we go for say called n, c . If they say 10, 2 this is n is the lot size. C is acceptance number.

What is the meaning of this in n, c is the lot size 10, I will count all the defective products. The number of defective ways 2 or less, I will accept the whole lot. If the number of defective is two or more, I will reject the whole lot. So, there is a, it is a kind of intuitive process there is mathematics behind it which follow by normal distribution. But it is very easy way but there is more variability still will be maintained.

When we go for statistical process control, when the process started then we are controlling the process parameters that should go for statistical process control like our control chart. But the design of experiment days before starting of the experiment, before manufacturing in at laboratory level, you can see what are the parameters that will affect the quality of the product. So, we can control the product so that you can control the variable so that you can improve the quality.

With help of design of experiments, we can maintain high level of quality. That is why people are interested in design of experiments. The base for this design of experiment is nothing but your anova, analysis of variance. That is why I am trying to connect the connection between the design of experiments and Anova.

(Refer Slide Time: 12:40)

Effect of Teaching Methodology		
Group 1 Black Board	Group 2 Case Presentation	Group 3 PPT
4	2	2
3	4	1
2	6	3

I am going to explain the concept of analysis of variance with the help of an example. The example is, there are three teaching methodology. The one way of teaching is with help of blackboard and other ways with help of case studies. Third way is PowerPoint presentations suppose. I want to know which teaching methodology is more effective or is there any difference? Is there any influence of teaching methodology on student performance?

Totally 9 student was taken in each group 3 students are allotted randomly. So what do you see here is the marks obtained by the students when they are studying blackboard group. They ask to study case presentation group in where the teacher is using only PowerPoint presentations. Whatever value which you are saying, what value you are seeing this is the marks. These are the marks. What is the null hypothesis here?

(Refer Slide Time: 13:43)

$$\begin{aligned}
 \bar{x}_1 &= \frac{4+3+2}{3} = 3 \\
 \bar{x}_2 &= \frac{2+4+6}{3} = 4 \\
 \bar{x}_3 &= \frac{2+1+3}{3} = 2 \\
 \bar{x} &= \frac{4+3+2+2+4+6+2+1+3}{9} = 3
 \end{aligned}$$

$H_0: \mu_1 = \mu_2 = \mu_3$
 $H_1: \mu_1 \neq \mu_2 \neq \mu_3$

$$SST = SST_{\text{treat}} + \underbrace{SS_{\text{error}}}_{SS_E} = \sum (x_i - \bar{x})^2$$

$$SST = SSB + SSE$$

$$\begin{aligned}
 SST &= (4-3)^2 + (3-3)^2 + (2-3)^2 + (2-3)^2 + (4-3)^2 + (6-3)^2 + (2-3)^2 + (1-3)^2 + (3-3)^2 \\
 &= 1 + 0 + 1 + 1 + 1 + 9 + 1 + 4 + 0 = 18 \quad 9-1 = 8
 \end{aligned}$$

$$\begin{aligned}
 SSB &= 3(3-3)^2 + 3(4-3)^2 + 3(2-3)^2 \\
 &= 0 + 3 + 3 = 6 \quad 3-1 = 2
 \end{aligned}$$

$$\begin{aligned}
 SSE &= (4-3)^2 + (3-3)^2 + (2-3)^2 + (2-4)^2 + (4-4)^2 + (6-4)^2 + (2-2)^2 + (1-2)^2 + (3-2)^2 \\
 &= 1 + 0 + 1 + 4 + 0 + 4 + 0 + 1 + 1 = 12 \quad 2 \times 2 = 4
 \end{aligned}$$

The null hypothesis is I am going to assume that null hypothesis is $H_0: \mu_1 = \mu_2 = \mu_3$, alternative hypothesis is $\mu_1 \neq \mu_2 \neq \mu_3$. If you see even this technique name is called Anova analysis of variance. But I am comparing mean. What are you going to do here with the help of the concept called variance and I am going to compare the mean of three populations.

Nothing to do with, I am not going to, compare the variance and comparing the mean of three populations. So, far the group one have taken the sample mean that is 3 for group 2 I have taken sample means it is 4, group 3 it is 2. Then I find overall the sample mean $4 + 3 + 2 + 2 + 4 + 6 + 2 + 1 + 3$ with 9 elements, equal to 3. What I am going to do is I am going to find out the overall variance. That I am going to call it as SST, total sum of square.

Here, why I am saying it as variance, see the variance formula, we know that the variance what is the formula? Variance is equal to $(\sum (X - \bar{X})^2) / (n - 1)$. This numerator that is $(\sum (X - \bar{X})^2)$ whole square that I am going to say sum of square. I am going to find out the overall variance that overall variance SST and I am going to group into two categories.

This variance is due to SS treatment plus variance, variance due to error. That error minus that their sum of square. So what I am writing SST is equal to total sum of square equal to some time there might be between columns $SSB + SSE$. So, obviously in the variance, due to treatment, that

is SSB is dominating we can see that the teaching methodology is influencing variable. First I am going to find out the overall variance.

For that variance I am going to find out only the numerator so that I am going to call it as SST. What is SST? How each element is away from overall mean. Overall mean is 3, so in the first column, the first element is 4. So, $4 - 3$ whole square the value in the second column is 3. $3 - 3$ the whole square + $2 - 3$ whole square upto 18. So this 18 is nothing but total sum of square. This 18 I am going to see how much variance is due to this teaching methodology.

So I am going to call it is SSB. Some books call it SS treatment. That is treatment sum of square. What is treatment sum of square? In the first column, there are three element is there. 3 minus the first column mean is 3. This 3, so the overall mean is 3. So $3 (3 - 3)^2$ + the second column so this 3 represents the number of samples in each column, this 3 represents in the second column there are 3 elements.

So, this 4 represents mean of the second column is 4 minus 3 is overall mean 3 whole square + for third column also there are three elements. The mean of the third column is 2. 2 minus 3 this 3 is overall mean whole square. So, this becomes 0, $4 - 1, 1$ square into $3 = 3$, $2 - 1 = 1$, is going to be 6. So this 6 is numerator of the variance that is SSB is 6. Then I will find out SSE. That is the variance error sum of square, the inherent variance.

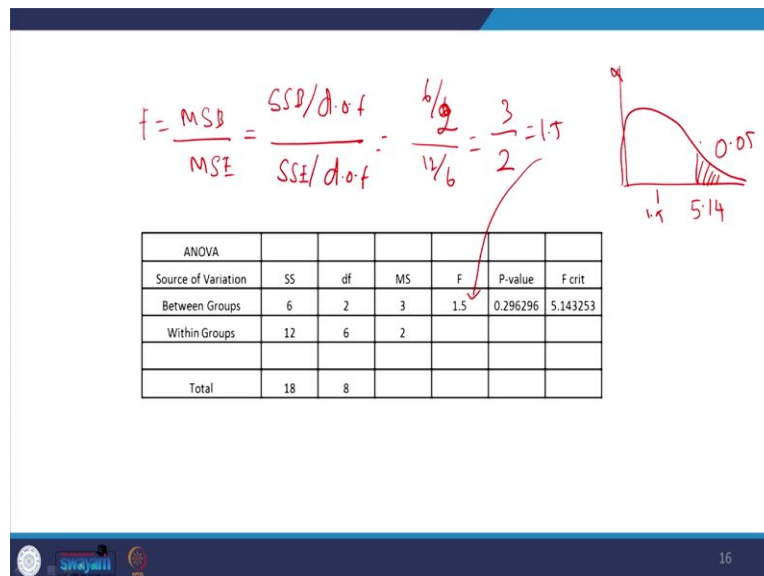
For inherent variance, when you look at the column 1 in the first element 4 the mean of the first column is 3. So, 4 minus 3 whole square + the second element is 3 minus the mean of the first column is this 3 this this this is a mean of first column. So, 4 minus 3 whole square + $(3 - 3)$ whole square + $(2 - 3)$ whole square. So this 4, this 4 represents mean of the second column at this represents this 2 represents the first element in the second column.

So, $2 - 4$ whole square + 4 minus 4 whole square + $6 - 4$ whole square. The third one is this 2 represents the mean of the third column. This 2 represent the first element in the third column $2 - 2$ whole square + $(1 - 2)$ whole square + $(3 - 2)$ the whole square. When you simplify that it is 12. So what happened the SST is divided into two categories one is variance due to treatment this

is Error variance or Individual variance. The second one is where to find out the degrees of freedom for SST.

What are the degrees of freedom? There are 9 elements. $9 - 1$ that is your degrees of freedom is 8. For SSB there 3 columns is there. So, 3 treatment. So, 3 minus 1 that is 3 minus 1 that is the degrees of freedom. For SSC, in the first column, there are three elements so the first column will have two degrees of freedom the second column the three elements. So, 3 minus 1, 2 degrees of freedom the third column also there are 3 elements. $3 - 1$, 2 degrees of freedom. So totally six degrees of freedom.

(Refer Slide Time: 19:32)



After that we have to find out f value. F value is nothing but MSB by MSE. That means what is MSB is mean square between columns. This is means error square. So what is this mean means, if you divide this SSB divided by the corresponding degrees of freedom we will get MSB. When you divide SSE divided by corresponding degrees of freedom we will get MSE. If we look at the previous one SSB, this is a two, degrees of freedom is 2. 6 divided by 2, SSE is how much chances?

That is 12. Degrees of freedom is 6, so 3 by 2 is 1.54 that 1.5 is nothing but this 1.5. Ok, Next what you have to do is, this is we can say your calculated value. You can refer your F table. In f table assume that alpha is equal to 5 %. You have to look at the, what is this value for example, if

you look at the table, it will give you 5.14 but your calculated value 1.5 is lying on the acceptance side table. This is F table so you have to accept null hypothesis.

This is an simple intuition for how this concept of Anova is working. In the next class we will do more theory behind this Anova, we will continue. Ok students in this class, we have seen how to find out the sample size for hypothesis testing. Then we have started the concept of Anova. For Anova I have taken one example then I have explained what is the SST, that is the total sum of square, treatment sum of square, error sum of square.

Then I have explained when should we go for Anova then I have solved one problem then, I will complete calculated F value with table F value then I have concluded the conclusion also. The next class, the same problem we will solve the help of python. Then, we will come then we'll go for post hoc analysis. Post hoc analysis is whenever you reject null hypothesis, we have to say which two pairs are different. So, that we will see in the next class, thank you very much.