

Applied Multivariate Statistical Modeling
Prof. J. Maiti
Department of Industrial Engineering and Management
Indian Institute of Technology, Kharagpur

Lecture - 14
Analysis of Variance (ANOVA)

(Refer Slide Time: 00:23)

Analysis of Variance (ANOVA)

No of population (L)	No of variables (P)	Hypothesis	Technique
L=1	P=1	H ₀ : $\mu = \mu_0$ H ₁ : $\mu \neq \mu_0$	t-test
	P ≥ 2	H ₀ : $\mu = \mu_0$ H ₁ : $\mu \neq \mu_0$	Hotelling's T ²
L=2	P=1	H ₀ : $\mu_1 = \mu_2$ H ₁ : $\mu_1 \neq \mu_2$	t-test
	P ≥ 2	H ₀ : $\mu_1 = \mu_2$ H ₁ : $\mu_1 \neq \mu_2$	Hotelling's T ²

Handwritten note: L=3 No of population

© CET
IIT KGP

NPTEL

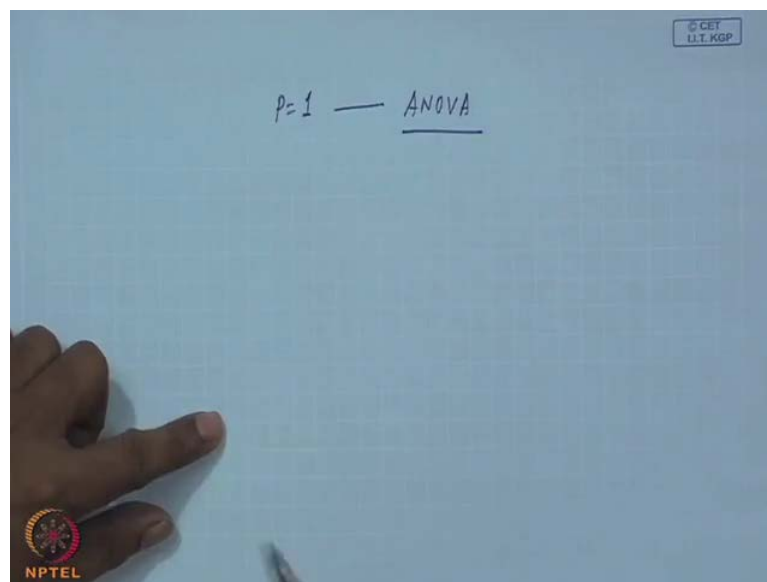
Good afternoon. Today, we will discuss Analysis of Variance popularly known as ANOVA. In last class, we have discussed about the difference between 2 population means, and in hypothesis testing we have covered the equality of population means from univariate point of view from multivariate point of view. So, if I just recapitulate this what we will find out that, one is number of population, then your number of variables, then hypothesis, then the techniques used, technique what is, what technique is used? If you say number of population 1 number of variable p and hypothesis as usual. Then the situation is single population l equal to 1, suppose number of variable is also one p equal to 1, what we do in hypothesis testing? We do $H_0: \mu = \mu_0$ and $H_1: \mu \neq \mu_0$.

Generally, we test this and we use for the small sample case t test, we assume that population is normal. Now, l can be 1, and p can be more than 2 all case also that is single population more than or equal to two variables. Your hypothesis will be again H_0

μ equal to μ_0 , here μ is a vector quantity and $H_1: \mu \neq \mu_0$. We have used Hotelling's T square for hypothesis testing again for small sample case.

Now, second issue we have discussed 1 equal to 2 population case, under this also we have already discussed p equal to one case. There our hypothesis was $\mu_1 = \mu_m$ and alternative hypothesis we have framed as $\mu_1 \neq \mu_m$ for at least one pair. We have used t test and for p greater than equal to 2 case, we have hypothesis like this that $\mu_1 = \mu_m$ here μ_1 and μ_m are vector quantities and $H_1: \mu_1 \neq \mu_m$ for at least one pair. We have also used Hotelling's T square, this is what from hypothesis testing point of view we have already completed.

(Refer Slide Time: 04:39)




What will happen when 1 is three or more? So, our discussion today for 1 greater than equal to 3 case, where 1 stand for population that is number of population. So, in this case we will first consider the univariate issue. That means we are interested in p equal to 1 case. When we test equality of several population means for one response variable the technique used is ANOVA analysis of variance. That mean what I said that suppose you case is p equal to 1 that is univariate case there are several groups or population, which is definitely three or more, in that case the first t test is not applicable. Simultaneously, if you want to test the difference you require to use a special technique known as analysis of variance.

(Refer Slide Time: 05:34)

Contents

- Conceptual model
- Assumptions
- Estimation of parameters
- Model adequacy tests
- Interpretation of results
- Reference



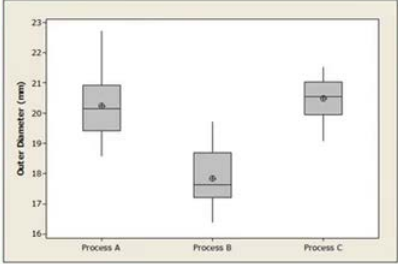
Dr J Maiti, IEM, IIT Kharagpur 2

So, today we will see in terms of ANOVA. For ANOVA these are the contents conceptual model, assumptions, estimation of parameters, model adequacy tests, interpretation of results, followed by references.


(Refer Slide Time: 05:57)

Conceptual model: An example

Sl. No.	Process A	Process B	Process C
1	20	17	20
2	21	17	20
3	20	19	21
4	21	17	20
5	23	16	21
6	19	19	21
7	20	18	22
8	19	18	19
9	19	18	22
10	20	20	20



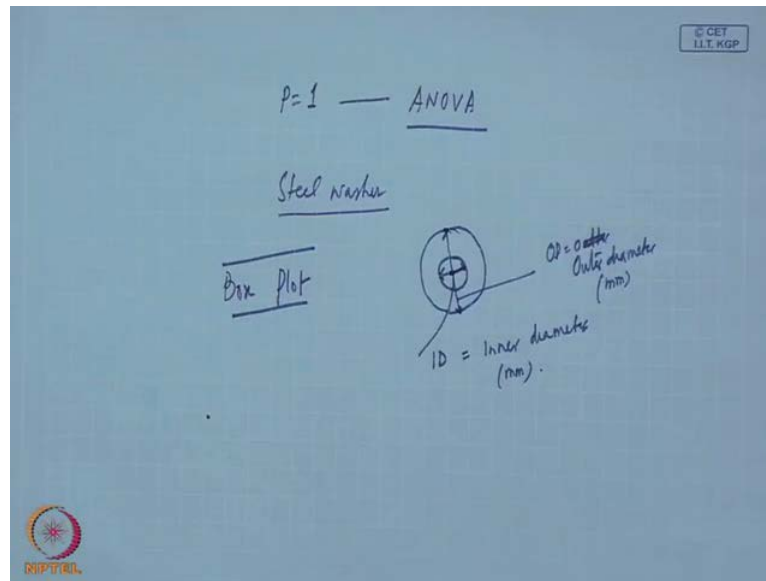
Are there mean differences in OD of the washers produced by any two of the three processes?



Dr J Maiti, IEM, IIT Kharagpur 3

What is this conceptual model? I will be describing with an example. You see this slide here we have three processes A B and C this three processes are producing let steel washers.

(Refer Slide Time: 06: 16)



So, our product is steel washers steel washer is the product, which can be in 2-dimensional figure. We can see like this, suppose this is the case and then this part or otherwise this inner circle part. If we consider this is I D inner diameter and if you consider the outer circle this will be O D, which is outer diameter outer. Let both are measured in millimetre is this case.

Here these values are in millimetre for outer diameter the problem is that you are producing through three different processes, but you are producing same thing that is steel washer and outer diameter is the quality variable and which is specified by the customers. Now, as a manufacturer of this steel washers what you require to know? You require to know that, whether all the processes are producing at the same level from quality point of view same level or not. The manufacturer is interested to test the differences in means of outer diameter for the three processes.

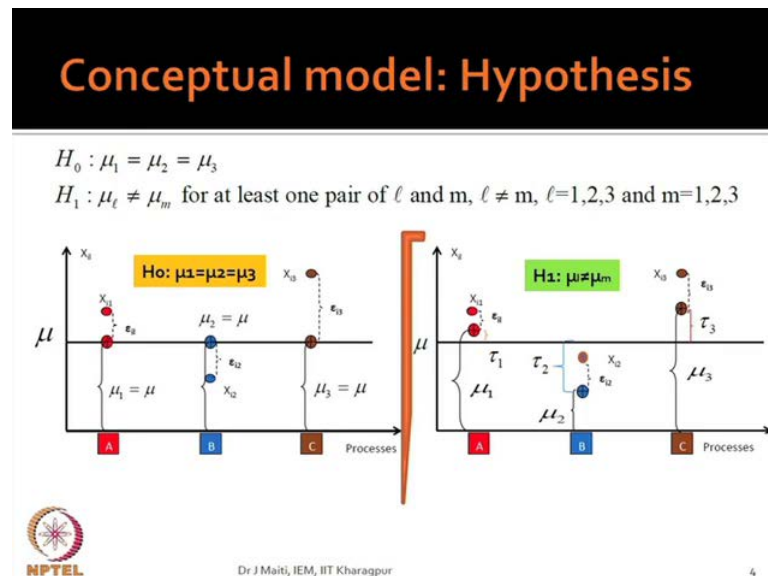
So, our objective here is are there mean differences in outer diameter of the washers produced by any two of the three processes, under such situation the most widely used graphical plot is box plot. Now, see what is box plot this figure shows? Box plot this box plot is for process A, this is for process B and process C. In the box plot there is one box and two whiskers this one is box the upper portion this straight line is right whisker and this is left whisker this box is determined by inter quarterly range I Q R. The horizontal

middle line here this is the median position from the data and the circle plus one that is the mean position.

So, if we see the three boxes here for process B and process a as well as process C. What we will find out that the box hardly differs in terms of variability? Except process C, but it is to be proven that whether this variability is different from A or B or not apparently it is different, but if you see the mean point the circle plus point this one is quite different. Apparently, it shows that there is mean difference you see this is the mean point for process a this is mean point for process B this one is mean point for process C. If I see the value it is 20.0, this one is around 18 and this one is again 20.0.

So, even if there is no difference between maybe process A means versus process C means, but there is chance that both process A and C means are different from B, which is basically from the graphical plot it resembles like this, but what you want you want a quantitative explanation of this. If we are able to do this then we are able to answer this question are there mean differences in O D of the washers produced by this.

(Refer Slide Time: 10:57)

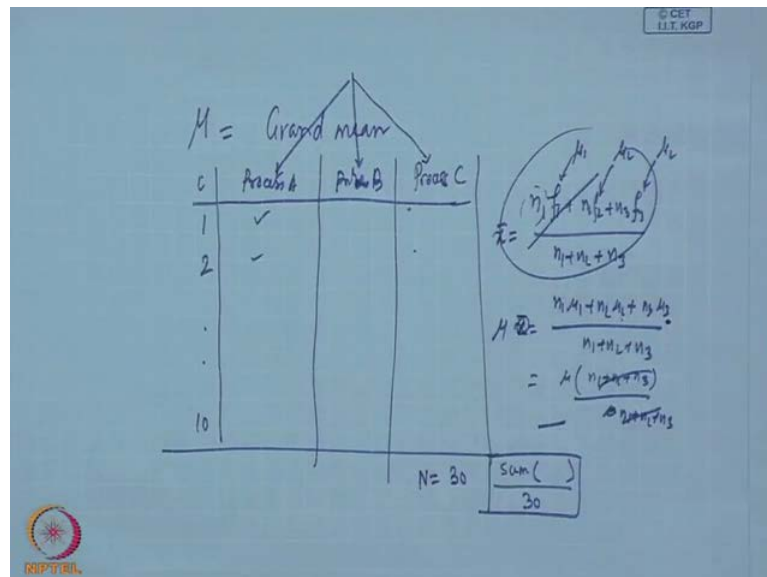


Then we will use that what I told as ANOVA uses two hypothesis 1 is null hypothesis. Alternative hypothesis what null hypothesis says null hypothesis says that there is no mean difference as you have taken 3 processes A B and C. So, we are saying that process A mean as μ_1 process B as μ_2 and process C as μ_3 . There is no difference means $\mu_1 = \mu_2 = \mu_3$ and our alternative hypothesis is atleast one of the

pairs. That means either $\mu_1 \mu_2$ pair or $\mu_1 \mu_3$ pair or $\mu_2 \mu_3$ pair is different that means are different.

So, pictorially the left hand side figure is your null hypothesis and here the right hand side is your alternative hypothesis if your situation is as the left hand side figure you say, what happened here there is one parameter μ . Please, see the left hand side this is μ here this μ is the grand mean. So, there is one parameter called μ , which is known as grand mean by grand mean what do I mean by grand mean we mean that that you have three processes here in this case. So, you have produced several items you measure the outer diameter in this case irrespective of the process you take everything as a whole and compute the means, ok?

(Refer Slide Time: 12:22)



This is what is grand mean the totality; for example, in this case we have seen that you will collect certain amount of data. We have collected ten data points sample size is 10 for process A process B and process C ten each. So, what will happen? Ultimately total data point if I say N that is 30 and you are considering every data points all sum. So, sum of all the data points divided by 30 that is your grand mean. If you collect 30 data parts 10 from each of the processes that will become your grand mean, ok?

Now, there is one another mean, which is known as process mean or in this case we will say the population mean. Because, we have assumed that each of the processes are different as if each of the processes are presenting a population process A is one

population process B is another process B is the other one μ_1 is the mean of process A μ_2 is the mean of process B and μ_3 is the mean of process C. When null hypothesis is true every $\mu_1 = \mu_2 = \mu_3$ then it will be equal to μ , because we are collecting. You are collecting some here ten samples from each of the population.

When null hypothesis is true, what will happen? You compute the process A mean process B mean process C mean you will also the grand mean we will be finding that they are equal, because it is obvious. Because, what is the group mean calculation do you know that I think all of you know $N_1 F_1 + N_2 F_2 + N_3 F_3$ by $N_1 + N_2 + N_3$. That is the group mean if you calculate group mean this is the formula, where for the first case frequency is F_1 second case frequency. This is basically $F_1, N_1, F_2, N_2, F_3, N_3$ and N_1 is the number of cases from the that is the I think we have used μ here $N_1 \mu + N_2 \mu + N_3 \mu$ is basically $\mu_1 \mu_2$ that will be better. So, I will write like this $N_1 \mu_1 + N_2 \mu_2 + N_3 \mu_3$ by $N_1 + N_2 + N_3$. So, this is the general expression we will not consider this one here.

So, if $\mu_1 = \mu_2 = \mu_3$ are same then what is happening here then it is $\mu \frac{N_1 + N_2 + N_3}{N_1 + N_2 + N_3}$. This will be cancelled out this is μ equal to μ , so grand mean will be like this. Now, let us see that what are the other parameters available here you see the left hand side again there is another parameter, which is ϵ_i this one this is the error quantity error in the sense.

If you take any observation X_{i1} from the first process from the first process, then if we consider that that $\mu_1 = \mu$ then μ_1 will come here the difference between the two that the observed value minus the mean value. That is ϵ_{i1} later on we will see this is basically error quantity. So, similarly for μ_2 also ϵ_{i2} for μ_3 , that is process C case also ϵ_{i3} you are getting.

Now, come to the right hand side in the right hand side, what is the figure resembles? That the μ_1, μ_2 and μ_3 are not same. So, they are not coinciding with μ and as a result what will happen? You will be getting one more parameter here that is known as τ . You see this is τ , this one is also τ , this is τ . So, τ_1 is this value, what is this? This is the process mean minus grand mean and τ_2 is again the process mean minus grand mean for process B τ_3 is process mean minus grand mean for process C So, one more parameter is coming into picture, so what we will do using all those things?

(Refer Slide Time: one 8:05)

Conceptual model: parameters

$$X_{i\ell} = \mu + \tau_{\ell} + \epsilon_{i\ell}$$

$$\tau_{\ell} = \mu_{\ell} - \mu$$

$$\epsilon_{i\ell} = X_{i\ell} - \mu_{\ell}$$


$$\sum_{\ell=1}^L \tau_{\ell} = 0, \text{ for equal sample size}$$

and

$$\sum_{\ell=1}^L n_{\ell} \tau_{\ell} = 0, \text{ for unequal sample size}$$

$$H_0 = \tau_{\ell} = 0, \text{ for all } \ell = 1, 2, \dots, L$$

$$H_1 = \tau_{\ell} \neq 0, \text{ for at least one } \ell.$$



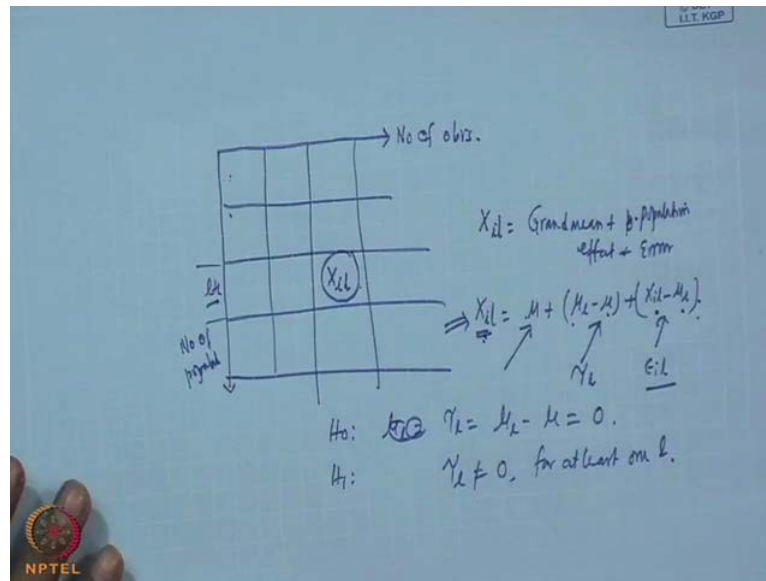
Dr J Maiti, IEM, IIT Kharagpur

5

We will redefine the ANOVA model and that is that any observation can be partitioned into three parts. That is $X_{i\ell}$ is the observation that is the i th observation on the ℓ th population to be collected μ is the grand mean τ_{ℓ} is the population effect and $\epsilon_{i\ell}$ is the error term, which is not captured either through grand mean or τ_{ℓ} or combiningly it is not captured basically. So, what is τ_{ℓ} ? τ_{ℓ} is the population effect. So, ℓ population effect is τ_{ℓ} , which is also known as treatment effect. So, this is $\mu_{\ell} - \mu$ that means the mean of ℓ th population minus grand mean and the other one $\epsilon_{i\ell}$ that is the error part is calculated like this.

Here we this right hand side there are two conditions give $\sum_{\ell=1}^L \tau_{\ell} = 0$ to $\sum_{\ell=1}^L n_{\ell} \tau_{\ell} = 0$. For equal sample size case means when you collect sample same sample size sample from the three populations or L populations what will happen? Here you will find out that their effect will be total effect will be 0. The n_{ℓ} is more than the other some cases some are more than the mean grand mean some are less than the grand mean, but the total effect will be 0. If you go for unequal sample size then your formula will be like this $\sum_{\ell=1}^L n_{\ell} \tau_{\ell} = 0$.

(Refer Slide Time: 20:02)




I think, I will, let me repeat this. What we are saying in ANOVA case we are saying that you have number of observations in this case here it is number of population. So, you may find out that several populations. Also, here will be several observations suppose I am saying that one observation is X_{il} that is the general observation that means i th observation on the l th population. This is my l population what we are saying further that X_{il} it can be decomposed as grand mean plus your population effect plus some error quantity. How do you get it we will get it like by this manner X_{il} equal to? I can write μ plus, let write like this $\mu_l - \mu$.

So, μ , μ cancelled out μ_l is there, but left hand side is X_{il} , so $X_{il} - \mu_l$ see μ and μ will be cancelled out $\mu_l - \mu_l$ will be cancelled out. So, X_{il} equal to X_{il} , so this is what is known as partitioning observation into three components here one is grand mean other one is the τ_l . That is the population effect and this one is ϵ_{il} . So, you are partitioning the observation values into three components here and each component is a parameter of ANOVA. That is from conceptual point of view that there are three parameters one is μ another one is τ_l and error component settings are there. If this is the case then your hypothesis can be changed to H_0 that is $\mu_l - \mu = 0$ and H_1 $\tau_l \neq 0$ for at least one l .

(Refer Slide Time: 22:51)

Assumptions

- Population variances are equal
- Errors are normally distributed
- Errors are iid



Dr J Maiti, IEM, IIT Kharagpur

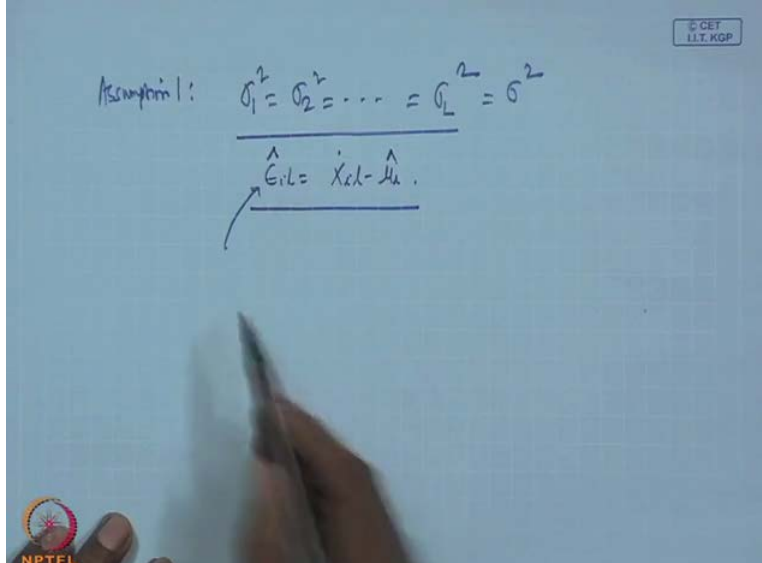
6

Like other multivariate models, ANOVA comes under univariate model like other multivariate model or like any model any statistical model. That ANOVA also requires certain assumptions to be satisfied these assumptions are populations variances are equal. So, you are sampling from 1 populations the population variances must be equal that means, what I mean to say sigma 1 square equal to sigma 2 square equal to sigma 1 square. This to be satisfied this is our first assumption, assumption one.


(Refer Slide Time: 23:27)

Assumption 1: $\sigma_1^2 = \sigma_2^2 = \dots = \sigma_L^2 = \sigma^2$

$\hat{\epsilon}_{ik} = X_{ik} - \hat{\mu}_k$



© CET
I.I.T. KGP



Assumption 2 is errors are normally distributed what is this error you have seen earlier that we have said that $\epsilon_i = X_i - \mu$. If you estimate this will be like this, so this error quantity is normally distributed errors are I I D mean independent and identically distributed. If I say error one is normally distributed then error N also normally distributed. That is identically distributed independent mean there is no correlation between the errors, ok?


So, normal distributions as well as I I D condition those things you know and how to test the normality we have also discussed earlier in some lectures. Now, I will first explain how to test the equality of population variances if the one is not satisfied this assumption 1? That is equality of population variances is not satisfied, what will happen? Then you cannot use the traditional ANOVA. It is there some other methods are there, but you have to go for different way of doing things, but here this assumption is vital for us.

(Refer Slide Time: 25:23)

Test of equality of population variances

- Bartlett's test

Hypothesis	$H_0 : \sigma_1^2 = \sigma_2^2 = \dots = \sigma_L^2$ $H_1 : \sigma_j^2 \neq \sigma_m^2, \text{ for at least pair of } (j, m).$
Statistic	$\chi_0^2 = 2.3026 \frac{q}{c}$ $q = (N - L) \log_{10} S_p^2 - \sum_{i=1}^L (n_i - 1) \log_{10} S_i^2$ $c = 1 + \frac{1}{3(L-1)} \left(\sum_{i=1}^L (n_i - 1)^{-1} - (N - L)^{-1} \right), \quad S_p^2 = \frac{\sum_{i=1}^L (n_i - 1) S_i^2}{N - L}$
Decision	Reject H_0 when $\chi_0^2 > \chi_{\alpha, L-1}^2$.


Dr J Maiti, IEM, IIT Kharagpur
7

So, for equality of population variances you will use Bartlett test. Now, see the format what we have written here hypothesis statistic and decision and that we have seen earlier. Now, what is our hypothesis? That there is no differences in the population variances and alternative hypothesis is at least one pair of variances are different. Now, here Bartlett proposed one statistic, which is $2.3026 \frac{q}{c}$. This is a statistic where q is $(N - L) \log_{10} S_p^2 - \sum_{i=1}^L (n_i - 1) \log_{10} S_i^2$. I hope you can recollect, what is S_p^2 ? And what is S_i^2 ?

This is pooled variance and this one is individual population variance. So, you require to compute first s_1^2 , s_2^2 , s_3^2 , s_4^2 like the all individual population variance values. Then also you require to compute the pooled one this pooled one is given in this equation $s_p^2 = \frac{1}{N-1} \sum_{i=1}^L (n_i - 1) s_i^2$. I think that we have already seen in last class.

(Refer Slide Time: 27:27)

Assumption: $\sigma_1^2 = \sigma_2^2 = \dots = \sigma_L^2 = \sigma^2$

$$\hat{\sigma}_{iL} = \frac{\sum_{k=1}^L (n_k - 1) s_k^2}{\sum_{k=1}^L (n_k - 1)}$$

$N = \sum_{k=1}^L n_k$

What we are saying that s_p^2 this is $\frac{1}{N-1} \sum_{i=1}^L (n_i - 1) s_i^2$ like this $\frac{1}{N-1} \sum_{i=1}^L (n_i - 1) s_i^2$ divided by $N-1$ plus $N-1$ to $N-1$ minus the how many populations are there that is L . So, this quantity is nothing but $\frac{1}{N-1} \sum_{i=1}^L (n_i - 1) s_i^2$ equal to $\frac{1}{N-1} \sum_{i=1}^L (n_i - 1) s_i^2$ divided by sum total of $N-1$ where L equal to 1 to capital L . So, this is your s_p^2 you have developed earlier also, if you know this then you are in a position to calculate q what is capital N here.

$N-1$ this capital N is nothing but this one capital N is 1 equal to 1 to capital L small N that means the total sample size from all the populations. Although, it is total sample size is not that meaningful mean from every population. You collected certain sample of size some sizes and that totality we are talking about.

Then you have to compute $C = \frac{1}{N-1} \sum_{i=1}^L (n_i - 1) s_i^2$ inverse minus $N-1$ inverse this is your C value. So, please go through this C there are basically C is another quantity, which is basically i and N values are defecting this C and this C will be used as the divisor. So, then chi square is $2.3026 q$ by C . So,


given data set you have to calculate q you have calculate C then you find out the chi square value chi square is 2.3026 q by C .

What does it mean, when you say chi square because this quantity follows chi square distribution. What will be the degrees of freedom? That you see it is clearly written here in the decision case that reject H_0 , when the chi square computed is greater than chi square $1 - \alpha$ for a particular alpha value. So, what do you mean then that 2.3026 follows chi square distribution with $l - 1$ degrees of freedom.

(Refer Slide Time: 30:55)

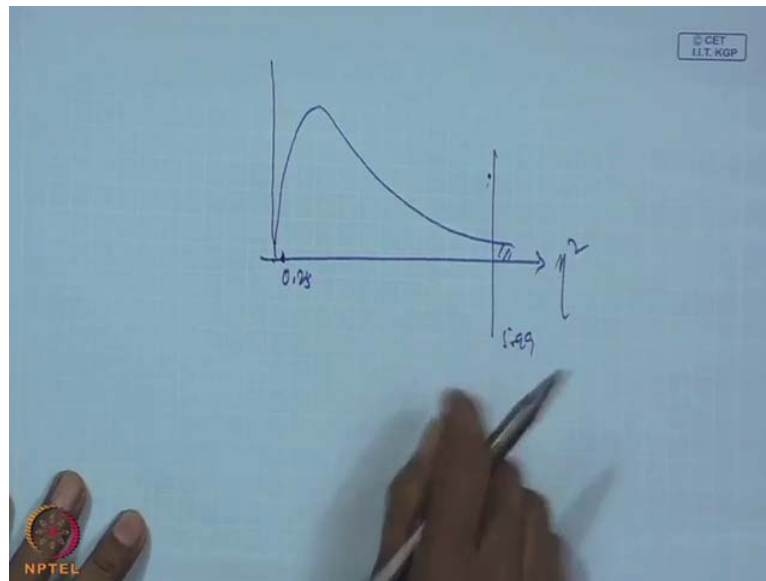
Bartlett's test

Process A	Process B	Process C		
20	17	20	s_1-sq	1.51
21	17	20	s_2-sq	1.43
20	19	21	s_3-sq	0.93
21	17	20		
23	16	21	S_p-sq	1.29
19	19	21	q	0.26
20	18	22	c	1.05
19	18	19	Chi-sq	0.25
19	18	22	Chi(2, 0.05)	5.99
20	20	20	Failed to reject H_0	


Dr J Maiti, IEM, IIT Kharagpur
8

I think the similar thing you have seen earlier also. I have been repeating, see the same problem. So, we have 30 observations 10 from each of the processes and what we have done here? We have calculated the variance for the first process that is 1.5 one variance for the second process that is 1.43 variance for the third process that is 0.93. Then you have computed chi square by that this one is chi square. So, you have computed first pool variance then q then C then chi square, which is 0.25. Now, we all know that chi square two with two degrees of freedom and alpha equal to 0.05. This value is 5.99 we have seen earlier, we have seen this one earlier also.

(Refer Slide Time: 32:08)



Now, the computed value is less than the tabulated value. So, what will be your decision you will accept null hypothesis or reject null hypothesis accept null hypothesis you fail to reject null hypothesis. So, the same principle this will be my chi square suppose this is your chi square distribution. Your alpha tabulated value here this is 5.9, but your value is how much 0.25, this is very close to 0. So, what does it signify? It signifies that there is no population variance differences all the population variances are equal same. So, we are fit for ANOVA case.

(Refer Slide Time: 32:51)

Decomposition of total sum of squares

Population	i = 1, 2, .., n	$\bar{x}_i = \frac{1}{n} \sum_{t=1}^n x_{it}$	Partitioning of observations (x_{it})
1	$x_{11}, x_{21}, \dots, x_{i1}, \dots, x_{n1}$	$\bar{x}_1 = \frac{1}{n} \sum_{t=1}^n x_{t1}$	$x_{it} = \bar{x} + (\bar{x}_i - \bar{x}) + (x_{it} - \bar{x}_i)$
2	$x_{12}, x_{22}, \dots, x_{i2}, \dots, x_{n2}$	$\bar{x}_2 = \frac{1}{n} \sum_{t=1}^n x_{t2}$	
⋮		⋮	
ℓ	$x_{1\ell}, x_{2\ell}, \dots, x_{i\ell}, \dots, x_{n\ell}$	$\bar{x}_\ell = \frac{1}{n} \sum_{t=1}^n x_{t\ell}$	
⋮		⋮	
L	$x_{1L}, x_{2L}, \dots, x_{iL}, \dots, x_{nL}$	$\bar{x}_L = \frac{1}{n} \sum_{t=1}^n x_{tL}$	
Grand mean		$\bar{x} = \frac{1}{nL} \sum_{i=1}^L \sum_{t=1}^n x_{it}$	

Dr J Maiti, IEM, IIT Kharagpur
9

In ANOVA, it will not go in the same manner like Hotelling t square or your other one that t test the way we have developed in ANOVA. It is a different ball game altogether, because here the primary concern is partitioning the observation. We have seen earlier, I told you that from the population point of view, when I am planning to collect some data let X_{il} . This is the observation to be collected from a population, where the grand mean is μ and population mean is μ_l . Then we have partition the observation like this you have seen this one see the left hand side and right hand side both are same. Now, you have collected data that means you have a fixed value X_{il} .

(Refer Slide Time: 33:11)

The image shows a handwritten derivation on a blue background. At the top right, there is a small box containing the text "© IIT KGP". The derivation starts with the equation $X_{il} = \mu + (\mu_l - \mu) + (X_{il} - \mu_l)$. Below this, it shows $X_{il} = \bar{x} + (\bar{x}_l - \bar{x}) + (X_{il} - \bar{x}_l)$. Then, it states $X_{il} - \bar{x} = (\bar{x}_l - \bar{x}) + (X_{il} - \bar{x}_l)$. The main part of the derivation is the summation of squares: $\sum_{i=1}^n (X_{il} - \bar{x})^2 = \sum_{i=1}^n (\bar{x}_l - \bar{x})^2 + 2 \sum_{i=1}^n (\bar{x}_l - \bar{x})(X_{il} - \bar{x}_l) + \sum_{i=1}^n (X_{il} - \bar{x}_l)^2$. The middle term is shown to be zero, leading to the final result: $\sum_{i=1}^n (X_{il} - \bar{x})^2 = \sum_{i=1}^n (\bar{x}_l - \bar{x})^2 + \sum_{i=1}^n (X_{il} - \bar{x}_l)^2$. There is a handwritten note "k-h population" with an arrow pointing to the first term of the final equation. In the bottom left corner, there is a logo for NPTEL.

So, you have to go by sample grand mean plus your population mean minus again grand mean plus X_{il} . That value minus population sample means not population sample mean. So, it is, now you do little manipulation $X_{il} - \bar{x} = \bar{x}_l - \bar{x} + X_{il} - \bar{x}_l$ square the term. If you square it $(X_{il} - \bar{x})^2 = (\bar{x}_l - \bar{x})^2 + 2(\bar{x}_l - \bar{x})(X_{il} - \bar{x}_l) + (X_{il} - \bar{x}_l)^2$ fine you take summation over i . So, i equal to one to N we are assuming that equal size samples are collected from the l populations. So, again i equal to one to N then this side will be i equal to one to N this side will be i equal to one to N .

Now, what will happen to this middle one? The middle one will become zero getting me, because of this quantity. So, when you write down sum total i equal to one to N some X_i this will be $N \bar{X}$ minus \bar{X} for N times that will be also $N \bar{X}$. So, this value will become zero, so your resultant quantity will be i equal to one to $N X_i$ bar minus \bar{X} square plus i equal to one to $N X_i$ bar minus \bar{X} square. Now, see that this one this is in the left hand side some quantity right hand side also, me quantity two squares is given here. This is for a particular population l population yes or no? Because, we have taken X_i everywhere l is there, but how many l populations are there? Capital l populations are there. So, that means your observation is spreaded over all l populations.

(Refer Slide Time: 37:33)

The image shows a handwritten derivation on a blue background. At the top left, it says "L-1 population" and "L". The main derivation is as follows:

$$\sum_{i=1}^n (x_{il} - \bar{x})^2 = \sum_{i=1}^n (\bar{x}_l - \bar{x})^2 + \sum_{i=1}^n (x_{il} - \bar{x}_l)^2$$

The second term in the sum is zero because $\sum_{i=1}^n (x_{il} - \bar{x}_l) = 0$. The derivation then shows the summation over all groups l :

$$\sum_{l=1}^L \sum_{i=1}^n (x_{il} - \bar{x})^2 = \sum_{l=1}^L \sum_{i=1}^n (\bar{x}_l - \bar{x})^2 + \sum_{l=1}^L \sum_{i=1}^n (x_{il} - \bar{x}_l)^2$$

The total sum of squares (SST) is then given by:

$$SST = \sum_{l=1}^L n (\bar{x}_l - \bar{x})^2 + \sum_{l=1}^L \sum_{i=1}^n (x_{il} - \bar{x}_l)^2$$

So, if take one more summation what will happen? What will you do? Then you will write down l equal to 1 to capital l sum total of i equal to 1 to $N X_i$ bar minus \bar{X} square equal to l equal to 1 to capital l i equal to 1 to $N X_i$ bar minus \bar{X} square plus sum total of l equal to 1 to capital l i equal to 1 to $N X_i$ bar minus \bar{X}_l square this is also square. Now, in this case is there any X_i bar and \bar{X} there is no i component. So, that mean the same component for N times, so what you can write like this l equal to one to capital l N will be there X_i bar minus \bar{X} square plus sum total l equal to 1 to capital l i equal to 1 to $N X_i$ bar minus \bar{X}_l square.

So, what will be the left hand side and the right hand side? If you see when we calculate the variance what we will do? We will subtract the mean value from all the observed values and then square it and then we divide it by the degrees of freedom here we have not divided it into degrees of freedom. Actually, what is happening here this $X_{i\ell}$ minus \bar{X} this was the variability this square. So, for all the observation case you have subtracted by the grand central average. So, this quantity is known as sum square total.


(Refer Slide Time: 39:57)

Decomposition of total sum of squares

$$x_{i\ell} - \bar{x} = \bar{x}_\ell - \bar{x} + x_{i\ell} - \bar{x}_\ell$$

$$\sum_{\ell=1}^L \sum_{i=1}^n (x_{i\ell} - \bar{x})^2 = n \sum_{\ell=1}^L (\bar{x}_\ell - \bar{x})^2 + \sum_{\ell=1}^L \sum_{i=1}^n (x_{i\ell} - \bar{x}_\ell)^2$$

SST	=	SSB	+	SSE
N-1	=	L-1	+	N-L


Dr J Maiti, IEM, IIT Kharagpur
30

You see here that $\sum_{\ell=1}^L \sum_{i=1}^n (x_{i\ell} - \bar{x})^2$ is the sum square total and we have already seen that other one $n \sum_{\ell=1}^L (\bar{x}_\ell - \bar{x})^2$ is the variability part from grand mean to the individual population means that sum we have taken you see this one what you have done here? \bar{x}_ℓ this is the population for ℓ th population sample average. \bar{x} is the grand sample average the difference between the two and then this observations you have squared. So, if there is no difference between the population means we assume that the sample average also will become same. It will be equal to the grand sample average, so then this quantity will become 0.

(Refer Slide Time: 41: 10)

Handwritten derivation on a whiteboard:

$$\sum_{i=1}^n (x_{il} - \bar{x})^2 = \sum_{i=1}^n (\bar{x}_l - \bar{x})^2 + \sum_{i=1}^n (x_{il} - \bar{x}_l)^2$$

$$\sum_{l=1}^L \sum_{i=1}^n (x_{il} - \bar{x})^2 = \sum_{l=1}^L \sum_{i=1}^n (\bar{x}_l - \bar{x})^2 + \sum_{l=1}^L \sum_{i=1}^n (x_{il} - \bar{x}_l)^2$$

$$\frac{SST}{N-1} = \frac{\sum_{l=1}^L n (\bar{x}_l - \bar{x})^2}{L-1} + \frac{\sum_{l=1}^L \sum_{i=1}^n (x_{il} - \bar{x}_l)^2}{N-L}$$

Additional notes on the board include "L-K population" and "L".

So, if there is any variability, this is because of the variability between the population. So, that is why this quantity is known as S s B and we have seen earlier also $X_{i1} - \bar{X}_1$ this is nothing but the error part, which is not explained by the which is not explained by S s b. So, this is S s e sum square error, now sum square total is equal to sum square between populations plus sum square error.

(Refer Slide Time: 41:50)

Handwritten notes on a whiteboard:

$$SST = SSB + SSE$$

$$\frac{N-1}{N} = \frac{L-1}{L} + \frac{N-L}{N} \leftarrow \text{DOF}$$

Diagram showing a data matrix with marginal means:

		\bar{x}_L
1		\bar{x}_1
2		\bar{x}_2
...		\vdots
...		\bar{x}_L
L		\bar{x}

Additional notes include $N = nL = \sum_{l=1}^L n_l$ and a circled $N-1$.

So, what is happening here? Then we are basically dividing the total variability in terms of sum square total equal to variability explained by the population plus variability not

explained by the population correct. So, this is the known as decomposition of decomposition, the total variability into between population variability and error variability.

So, now we have to see the degree of freedom part degrees of freedom. So, when we compute $S_s t$ what we require basically we require \bar{X} . So, there are total N observations and we have sacrificed one to calculate this one \bar{X} . So, degree of freedom for $S_s t$ is N minus capital N minus one where capital N is N into l when you sample from l population and sample size is equal. If sample size is unequal then this will be your l equal to one to capital N n l getting me. So, this is for equation for the equal sample size case this is the equation for computing capital N for the unequal sample size case.

Now, come to the between sum square, now in between sum square how many levels you have if you see the computation what we will be finding out? We will be finding out the population 1 to population l and you will be computing here that \bar{X}_1 . So, here \bar{X}_1 \bar{X}_2 like \bar{X}_l . Then you find out the \bar{X} and if you see the computation of $S_s b$ you are finding out that this is nothing but \bar{X}_1 minus \bar{X} square. Then multiplied by N and that summation essentially you are using \bar{X}_1 . So, as there are there are l populations and there is relationship also with the individual population mean this sum is grand mean. So, one degree you are losing here and as a result you have total l data points and one degree is lost by computing \bar{X} grand mean.

So, plus N minus l , because this is the rule that the total variability will also be decomposed into the component degrees of freedom. So, total degrees of freedom is N minus one and your $S_s b$'s degrees of freedom l minus 1 $S_s t$ degrees of freedom N minus l . So, if you make a sum l minus 1 plus N minus l it is nothing but N minus 1 right hand side N minus l plus l minus 1 sum total will be N minus 1 and which is equal to this keep in mind, this one. Yes, you are finding out N a, what you are using you are using l populations. So, you have l mean values.

(Refer Slide Time: 46:09)

Handwritten derivation on a blue background:

$$\Rightarrow x_{ki} - \bar{x} = (\bar{x}_k - \bar{x}) + (x_{ki} - \bar{x}_k)$$

$$\sum_{k=1}^L \sum_{i=1}^n (x_{ki} - \bar{x})^2 = \sum_{k=1}^L \sum_{i=1}^n (\bar{x}_k - \bar{x})^2 + 2 \sum_{k=1}^L (\bar{x}_k - \bar{x}) \sum_{i=1}^n (x_{ki} - \bar{x}_k) + \sum_{k=1}^L \sum_{i=1}^n (x_{ki} - \bar{x}_k)^2$$

$$= \sum_{k=1}^L n (\bar{x}_k - \bar{x})^2 + \sum_{k=1}^L \sum_{i=1}^n (x_{ki} - \bar{x}_k)^2$$

Below this, the derivation is repeated with more annotations:

$$\sum_{k=1}^L \sum_{i=1}^n (x_{ki} - \bar{x})^2 = \sum_{k=1}^L \sum_{i=1}^n (\bar{x}_k - \bar{x})^2 + \sum_{k=1}^L \sum_{i=1}^n (x_{ki} - \bar{x}_k)^2$$

$$\frac{N=nL}{N-1} \frac{SST}{L-1} = \frac{\sum_{k=1}^L n (\bar{x}_k - \bar{x})^2}{L-1} + \frac{\sum_{k=1}^L \sum_{i=1}^n (x_{ki} - \bar{x}_k)^2}{nL-L}$$

At the bottom, there is a diagram showing a horizontal line with an arrow pointing to the right, labeled 'L'. Below it, a box contains the symbols \bar{x}_k and \bar{x} .

Second one what is happening N into L this is the total observations. So, we are talking about N equal to small N into L this is the total observation this side is the total observation is this, but you have already computed X bar. So, this one this portion see how many X I you have computed? Bar, L populations. So, total is N n minus 1, so there is 1 X 1 bar to compute this one you require to know all those X 1 bars. So, already 1 degrees of freedom you have lost you have capital N degrees of observations 1 is lost. So, N minus 1, so this is what is the decompositioning the degree of freedom. Correct?

(Refer Slide Time: 47:24)

Hypothesis testing

Sources of variation	Sums square (SS)	Degrees of freedom	Mean square (MS)	F	Reject H_0
Population (treatment)	SSB	L-1	$MSB = \frac{SSB}{L-1}$	$MSE = \frac{SSE}{nL-L}$	$F > F_{L-1, nL-L}^{(\alpha)}$
Error (random component)	SSE	nL-L	$F = \frac{MSB}{MSE}$		
Total	SST	nL-1			

Dr. J. Maiti, IEM, IIT Kharagpur
12

This total concept whatever we have discussed, so far can be seen in a table, which is known as ANOVA table whenever you go through any textbook on ANOVA or any software you use for ANOVA what will happen you will be finding out that one table will be formulated. This table is known as ANOVA table which is very popular table later on even in regression other cases also we will be using ANOVA table.

This ANOVA table for the problem we have considered. Now, there are few items one is sources of variation then the sum square what we have calculated $S_s t$ and $S_s B$ all those things degrees of freedom also. We have seen then you have to compute mean square you have to compute f value then based on f value we will be accepting or rejecting the null hypothesis. So, we have considered 1 population no other factors we have considered in this present case sources of variation is the different populations. Apart from this you cannot nullify the random effects. So, random effect is coming under error, so another source of variability is error and this population and error these two sources are making the total variability.

So, as a result your table looks like this population error and total then $S_s b$, $S_s e$ and $S_s t$. You have already seen that, what is the degrees of freedom case l minus one for $S_s b$ for $S_s e$ N minus l and for $S_s t$ N minus 1, this N multiplied by capital l . This is the capital N , what I have discussed earlier that is the capital N this capital N this is N into 1, but if you take unequal sample size this is the different one.

Now, what is $M S B$? $M S B$ is mean square between populations. This is nothing but sum square between population divided by its degrees of freedom. Then what will be your $M_s e$? $M S e$ is $S_s e$ by its degrees of freedom. So, this $M S e$ will come here you write like this $m S B$ is this $m S e$ here under error you write under error $M S e$. So, I am writing again what will happen here? Population, then error, then total these are the sources of variation sources we are computing sum square.

(Refer Slide Time: 50:35)

Source	SS	DF	MS	F
Population	SSB	$L-1$	$\frac{SSB}{L-1} = MSB$	$\frac{MSB}{MSE}$
Error	SSE	$N-L$	$\frac{SSE}{N-L} = MSE$	
Total	SST	$N-1$		

$F(\alpha)_{L-1, N-L}$

$F > F(\alpha)_{L-1, N-L}$

$N = nL$


Now, it is sum square between populations sum square error and sum square total what is our degree of freedom? Degree of freedom it will be 1 minus one and this will be your N minus 1 and this will be N minus one, where N is nothing but capital 1 into small N and then we are talking about M S mean square, which is S s b by N minus 1 minus one. This will be m S s, S e by N minus 1 what is this S s B by 1 minus one is m S B S s e by 1 N minus 1 is M e.

Now, you are finding out a statistics called f, which is m S B divided by M S e, why it is f distributed? Ratio of 2 chi square variable, because S s B is the square this is sum square and sum square this is f distributed. What will be the degrees of what will be the degrees of numerator and denominator degrees of freedom for f here definitely what will be there in numerator S s b by degrees of freedom 1 minus 1. So, we will be comparing this with 1 minus one and what is the denominator degrees of freedom N minus 1. You will find out certain value of alpha. If your computed f that f computed here is greater than f 1 minus one N minus 1 into alpha. You will reject the null hypothesis, so once you reject null hypothesis you will accept the alternative hypothesis, that is the issue.

(Refer Slide Time: 53: 11)

Hypothesis testing

Sources of variation	Sums square (SS)	Degrees of freedom	Mean square (MS)	F	Reject H_0
Population (treatment)	43.415	2	21.708	22.82	p = 0.000
Error (random component)	25.686	27	0.951		
Total	69.101	29			

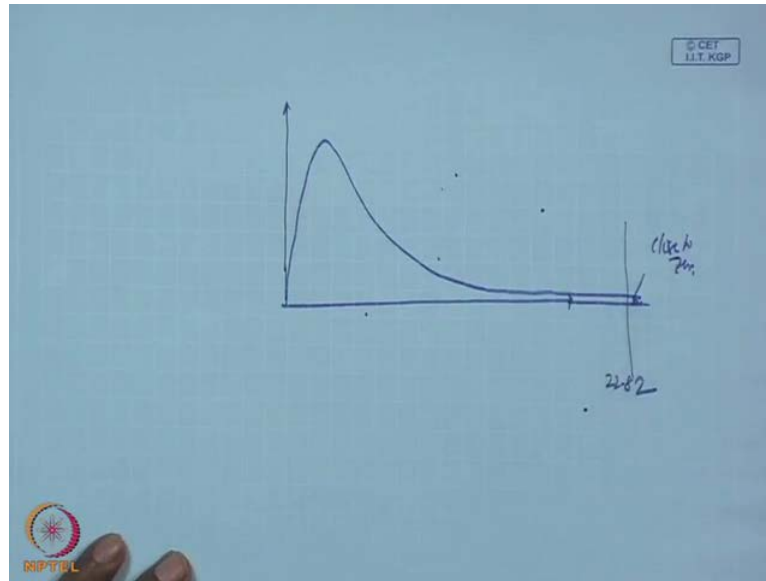


Dr J Maiti, IEM, IIT Kharagpur

33

The problem in the same problem we have done the hypothesis testing and population that variability sum square between population is 43.415. Error S s e is 25.686 and total is 69.101 degrees of freedom, because there are three processes. So, three minus one that is two three minus one is two and there are thirty total observations that minus one is 29 and you the difference between 29 and two is 27 that will be definitely the error's degree of freedom then mean square is 43.415 by 2. That is 21.708 and for error that is again 25.686 by 27, which is 0.95 one you see the mean square value is very less it is, so less because there is effect. So, then f is M s e by S s, M S, M S B by M s e. This is 21.708 is M S B and 0.951 is m S e and that ratio is 22.82 and which is far away from 0.

(Refer Slide Time: 54:46)



If you see the f table you will find the probability is almost close to 0. That mean, if I say like this is my f distribution that mean you are somewhere here this one, which is your 22.82. This one is close to 0, so there is difference this is what is tested through tested through ANOVA first level test is this there is difference in population our this one the means, population means that is what you are testing.

Now, here when you test this f test when you are doing here your, what is your null hypothesis? Null hypothesis, no population, no differences in population means correct. What is alternative hypothesis at least one pair is different. So, when you complete this one and if you find that you are rejecting H_0 . It simply says that there is population difference in terms of population means, but you do not know which they are, so we have to know, which pair are different? So, next class I will explain this which pair is different and other things.