

**INDIAN INSTITUTE
OF
TECHNOLOGY
KHARAGPUR**

**NPTEL
National Programme
on
Technology Enhanced Learning**

Applied Multivariate Statistical Modeling

**Prof. J. Maiti
Department of Industrial Engineering and Management
IIT Kharagpur**

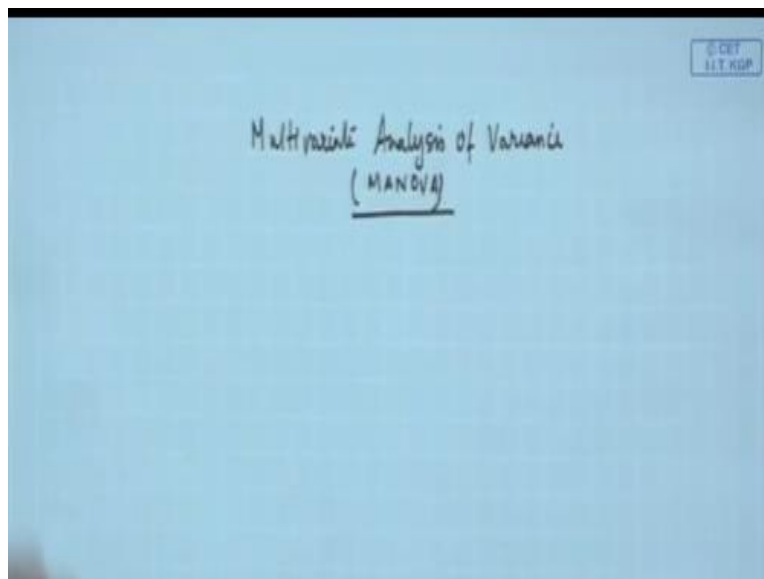
Lecture – 16

Topic

**Multivariate Analysis of Variance
(MANOVA)**

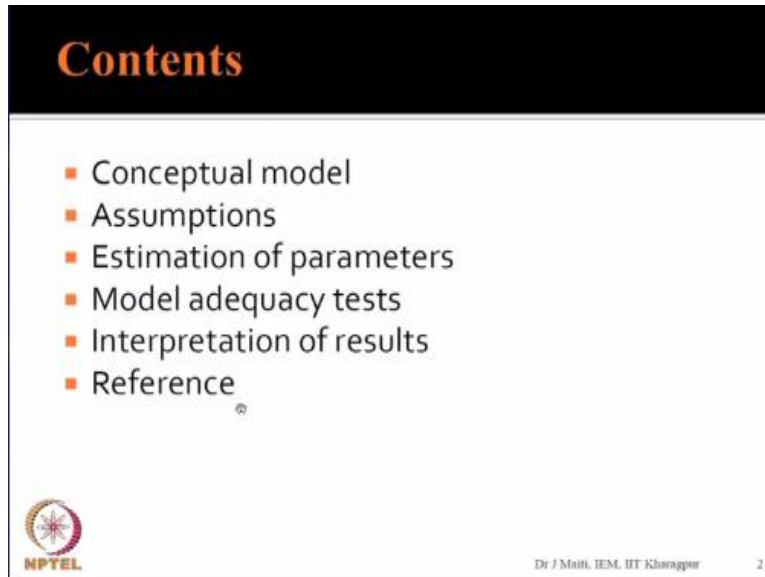
Good morning today, we will discuss Multivariate Analysis of Variance.

(Refer Slide Time: 00:23)




Multivariate analysis of variance, which is popularly known as MANOVA today's contents.

(Refer Slide Time: 00:45)



Contents

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

 NPTEL

Dr J Mathi, IEM, IIT Kharagpur 2

Are conceptual model for MANOVA, assumptions for MANOVA, modeling estimation of parameters, model adequacy tests, interpretation of results, and references. We will see that how much is possible to complete today and the remaining portion, we will be completing in the next class.

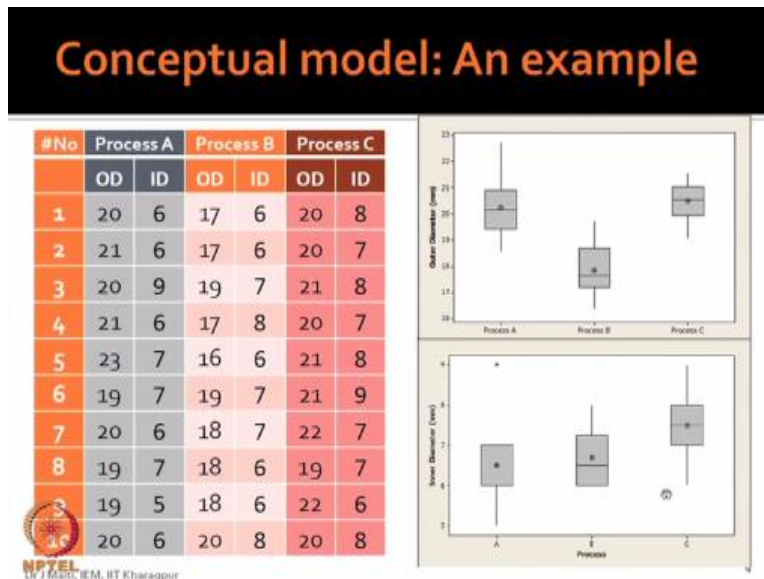
(Refer Slide Time: 01:10)

Conceptual model			
#populations (l)	#variables (p)	Hypothesis	Technique used
$l = 2$	$p = 1$	$H_0: \mu_1 = \mu_2$ $H_1: \mu_1 \neq \mu_2$	t-test
	$p \geq 2$	$H_0: \mu_1 = \mu_2$ $H_1: \mu_1 \neq \mu_2$	Hotelling's T-square
$l \geq 2$	$p = 1$	$H_0: \mu_1 = \mu_2 = \dots = \mu_L$ $H_1: \text{at least one pair } (\mu_l = \mu_m) \text{ is not equal}$	ANOVA
	$p \geq 2$	$H_0: \mu_1 = \mu_2 = \dots = \mu_L$ $H_1: \text{at least one pair } (\mu_l = \mu_m) \text{ is not equal}$	MANOVA

You see this slide last class I had shown you that when $l = 2$ where l stands for number of population and p is 1 then we have used t-test. That is the difference between two population mean that is described through has $H_0, \mu_1 = \mu_2$ and $H_1 \mu_1 \neq \mu_2$. For the $p \geq 2$ case that is the multivariate case here also this $H_0, \mu_1 = \mu_2$ and $\mu_0 = \mu_2$, but this μ_1 and μ_2 are in the vector domain. When $p = 2$ that is the two cross one vector and we have used Hotelling's T-square. Last class I have told you that, when $l \geq 2$ that is three or more for one variable case you will be using ANOVA.

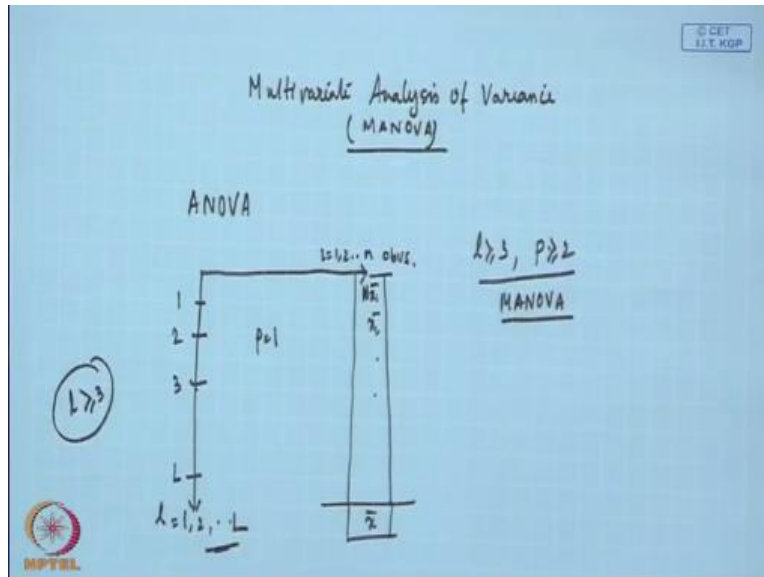
And we have discussed one way ANOVA, two way ANOVA, three way ANOVA and multi way ANOVA concept. Now, if your number of population is more than two that is three or more and as well as number of variables are more than two or more, in that case what will happen? You will find out that you require to use MANOVA not ANOVA.

(Refer Slide Time: 02:39)



So, essentially what we have discussed then when we will go for ANOVA.

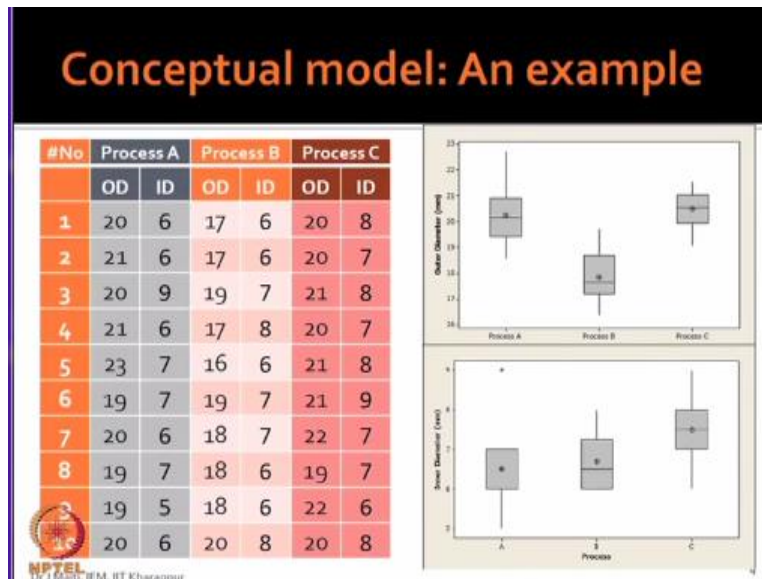
(Refer Slide Time: 02:44)



ANOVA is you have L number of population L can be 1, 2 L . That is 1, 2, 3 like this, this is the 1, 2, 3 and you will collect i^{th} number $i = 1, 2, \dots, n$ number of observations and you are interested to see the difference in population means, where population is determined by l_1, l_2 like 1 to L , that is the case everywhere the mean value you will find out here the mean value, you will find out mean one we can say $\bar{x}_1 \bar{x}_2$ like this. Then finally, \bar{x} that what we have seen last class now, in case in case of ANOVA that $L \geq 3$, then what you mean to say that if there is three or more population with 1 variable $p = 1$ you are going for ANOVA.

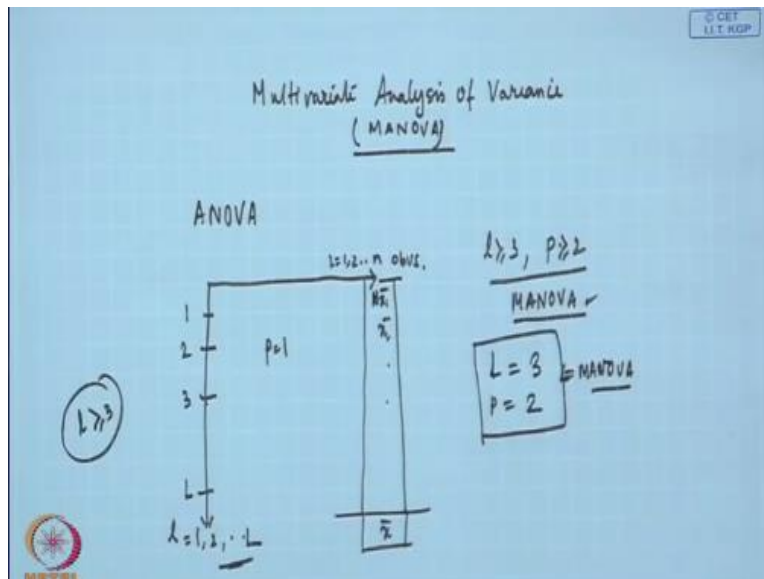
And when $L \geq 3$, but p is two or more then you will go for MANOVA. Your hypothesis, what you will propose we will see later on.

(Refer Slide Time: 04:17)



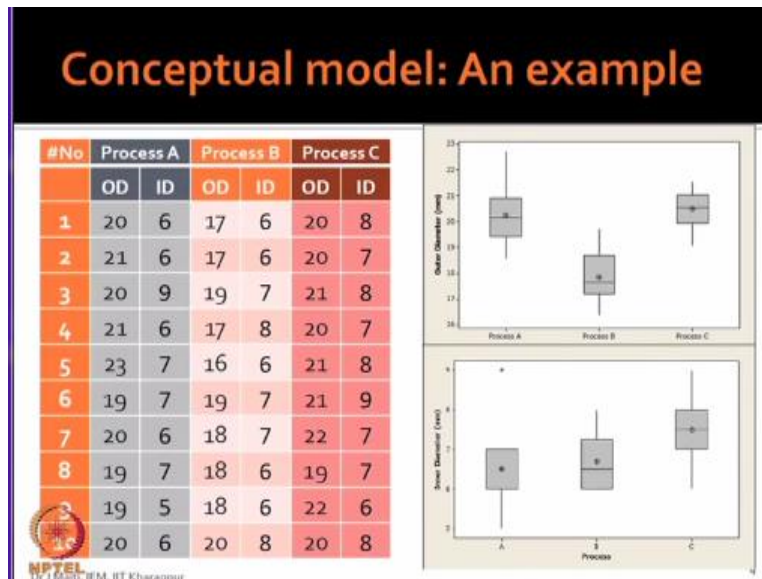
But before that let us see one example this is an example last class in ANOVA we have seen that process A, process B, process C producing steel washers with certain quality characteristics that is outer diameter. So, with 1 quality variables, now here we are adding one more quality variable that is inner diameter. So, that means the steel washers produced by the three processes are measured in terms of their two quality characteristics that is outer diameter and inner diameter. So, here $p = 2$ and there are three processes process A, process B and process C.

(Refer Slide Time: 05:05)



So, our case example case is $L = 3$ and $p = 2$. So, this a case for MANOVA, specific case for MANOVA, what we have discussed here okay I told you in last class also one of the useful way of looking into data is the box plot and if you see the box plot for the two different variables.

(Refer Slide Time: 05:30)

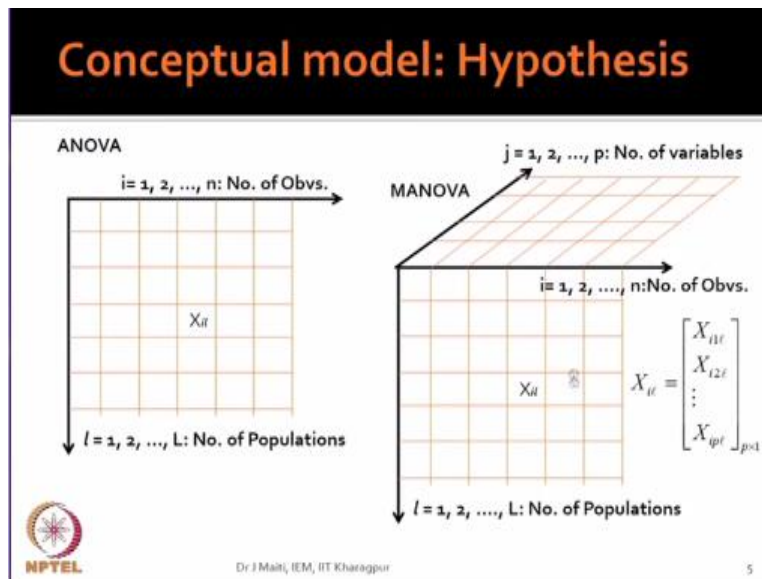


Inner and outer diameter for the three different processes process A, process B, process C. You are seen in last class for outer diameter the mean differences are quite visible from the box plot and if you see the lower figure where it is the inner diameters are plotted in terms of box and which curves so, you find out the mean value for process A, that mean the mean inner diameter is this one mean inner diameter for process B is this mean inner diameter for process C is this one.

So, apparently if we want to say about the differences in means between the three processes are process A and B means are not different, but process C is different from process A and B. If you see the outer diameter case here process A and C's mean differences are not all significant may be, but process B it is different from both A and C. So, there are two types of pictures, now from outer diameter point of view you are saying that and also, we have seen in ANOVA that process B is different than processes A and C, but here if we add inner diameter, what is happening here is that process C perhaps will differ in terms of their mean values of inner diameter from A and B.

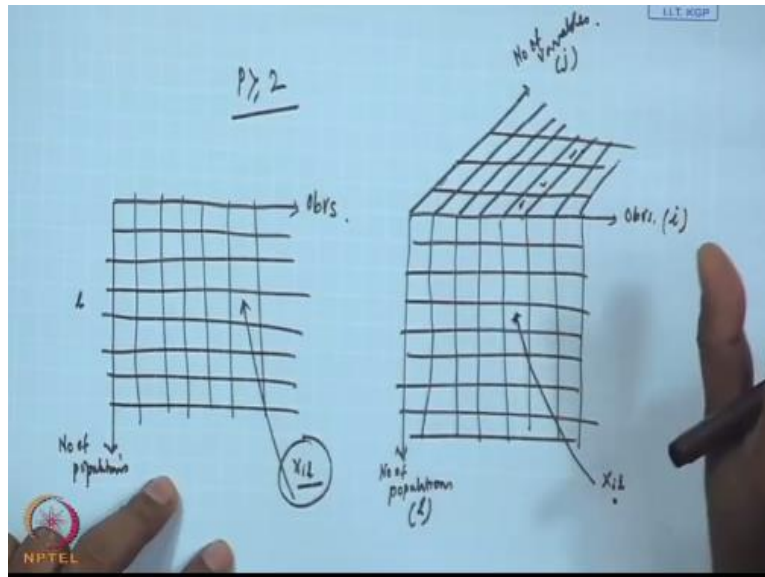
Now, we want to see collectively what is the difference in mean vector where this vector will be determined by mean diameter for outer diameter as well for mean inner diameter okay so, collectively are the two process, three processes different or not that is what we will be testing through MANOVA.

(Refer Slide Time: 07:37)



So, we require certain notations to fully understand the application of ANOVA as well as MANOVA may be in case of MANOVA there is one more dimension is added as number of variables.

(Refer Slide Time: 07:58)



Will be more than equal to 2. So $p \geq 2$, so as a result you see in ANOVA what we said we said like this that ANOVA case here it is number of populations this axis this axis is observations and for different population you have obtained certain observations and somewhere, there is one observation, which is X_{il} i^{th} observation on the l^{th} population and it is a scalar quantity. Now, in case of MANOVA what will happen you will find out that one this side is number of observation. Let it be population in the same manner number of populations and this side is observation, that is number of observations.

Then we will we will add one more dimension, which is number of variables so, if we denote number of population in terms of l number of observation in terms of i and number of variables in terms of j . Then what will happen your general structure will be like this is the data structure basically getting me, what we learn from any observation. Suppose, if I say this is my X_{il} getting me, this is our X_{il} then X_{il} is no longer a scalar quantity in ANOVA. This is a scalar quantity in MANOVA, it is not a scalar quantity, because for this cell you see if I go 1, 2, 3 depending on the number of variables X_{il} will have more number of values. So, we can say here that X_{il}

(Refer Slide Time: 10:27)

$X_{iL} = \begin{bmatrix} X_{i1} \\ X_{i2} \\ \vdots \\ X_{ip} \end{bmatrix}_{p \times 1}$

$\mu_L = \begin{bmatrix} \mu_{11} \\ \mu_{12} \\ \vdots \\ \mu_{1p} \end{bmatrix}_{p \times 1}$

$\Sigma_L = \begin{bmatrix} & & & \\ & & & \\ & & & \\ & & & \end{bmatrix}_{p \times p}$

$l = 1, 2, \dots, L$

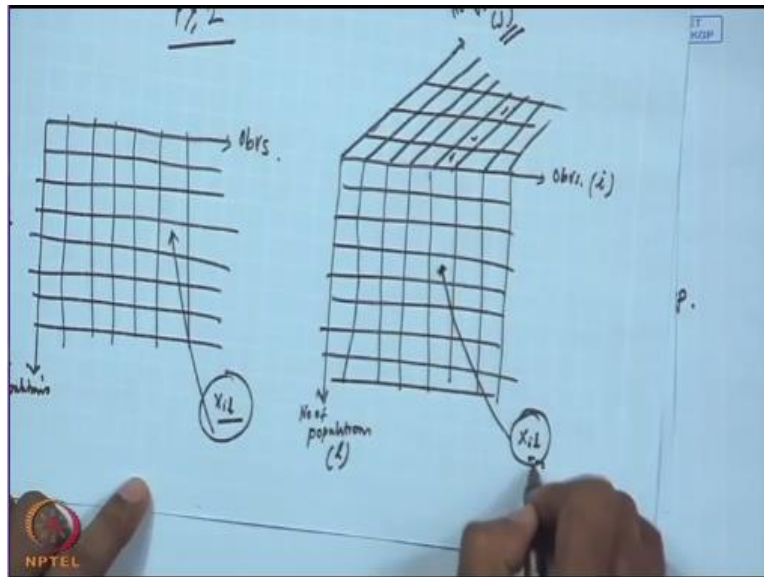
$\Sigma_1, \Sigma_2, \dots, \Sigma_L$

$\mu_1, \mu_2, \dots, \mu_L$

Is no longer a scalar quantity, it is a vector what we will write here X_{i11} , X_{i12} , so like this X_{i1} how many variables are there p variables are there $p \times 1$ and there is the complexity, because one more dimension is added and your total work is now in three dimensional case, it is not a two dimensional issue if this the case, so this is the general observation okay this is my general observation as you have p variables so, you also have p mean values, so now I am writing that the mean vector for population 1 there will be p mean values μ_{11} μ_{12} like. This μ_{1p} $p \times 1$ and as there are p variables again there will be one covariance matrix.

If I write like this $S_L \Sigma_1$ instead of s , we will be writing when we take the sample now, we will be writing like this. Then what will happen this will also be $p \times p$ matrix p variables are there your l varies from 1 to L . So, then there will be Σ_1, Σ_2 like this Σ_L . So, there will be similarly μ_1, μ_2, μ_L for mathematical simplicity we will be using this term.

(Refer Slide Time: 12:47)



Although, this a vector basically we will be using this term like this X_{iL} without bringing the variable part the variable part is implicit here, because X_{iL} is nothing but this one.

(Refer Slide Time: 13:05)

$X_{kL} = \begin{bmatrix} x_{k1} \\ x_{k2} \\ \vdots \\ x_{kp} \end{bmatrix}_{p \times 1}$ $\mu_k = \begin{bmatrix} \mu_{k1} \\ \mu_{k2} \\ \vdots \\ \mu_{kp} \end{bmatrix}_{p \times 1}$ $\Sigma_k = \begin{bmatrix} \end{bmatrix}_{p \times p}$

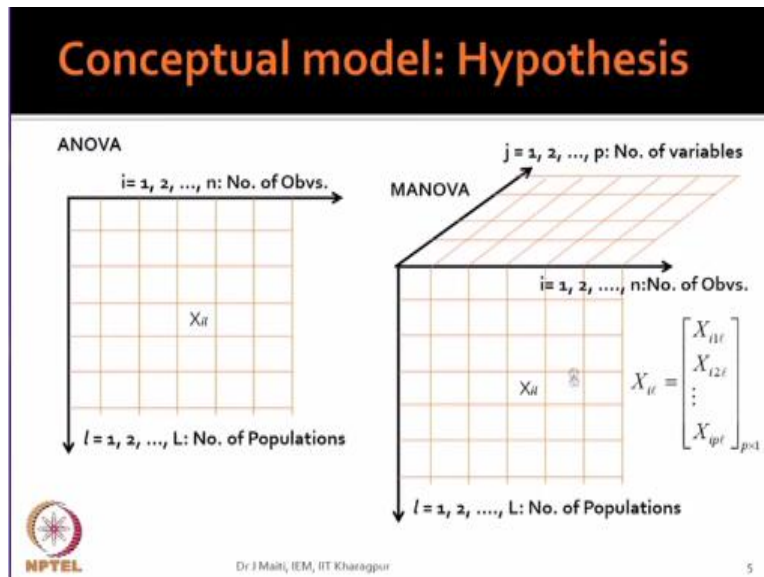
$k = 1, 2, \dots, L$

$\mu_1, \mu_2, \dots, \mu_L$

$\Sigma_1, \Sigma_2, \dots, \Sigma_L$

This is the general observation okay so, you see this pictorially given in this figure.

(Refer Slide Time: 13:13)



I think it is very clear. Now, for all you that ANOVA where all we have different populations and different observations and for MANOVA, it is not population observation and variables and this general observation is X_{iL} which is this one you have to keep in mind this vector part.


(Refer Slide Time: 13:42)

Conceptual model: Hypothesis

$H_0 : \mu_1 = \mu_2 = \dots = \mu_L$
 $H_1 : \mu_\ell \neq \mu_m$ for at least one pair of ℓ and m , $\ell \neq m$, $\ell=1,2,\dots,L$ and $m=1,2,\dots,L$

$$X_{i\ell} = \mu + (\mu_\ell - \mu) + (X_{i\ell} - \mu_\ell)$$
$$= \mu + \tau_\ell + \varepsilon_{i\ell}$$
$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix} \quad \tau_\ell = \begin{bmatrix} \tau_{1\ell} \\ \tau_{2\ell} \\ \vdots \\ \tau_{p\ell} \end{bmatrix}$$
$$H_0 : \tau_1 = \tau_2 = \dots = \tau_L = 0$$
$$H_1 : \mu_\ell \neq 0$$

\otimes

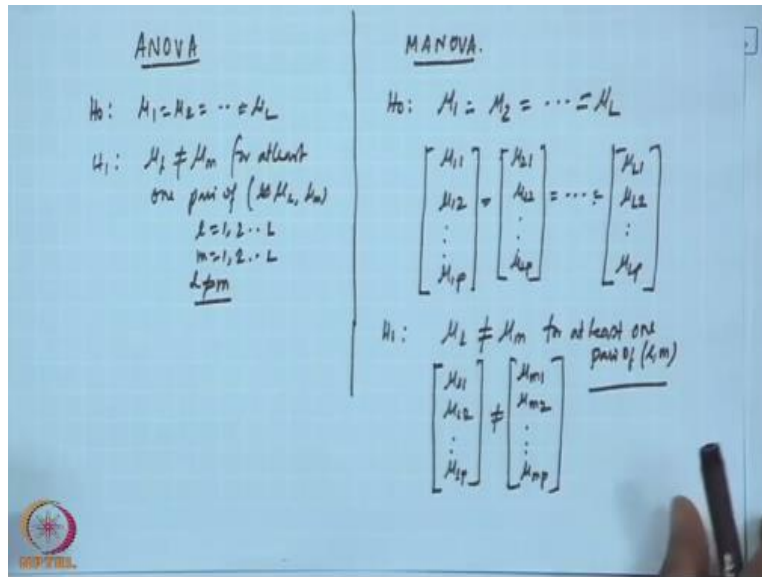


Dr J Mohi, IEM, IIT Kharagpur

6

Then MANOVA assume something and MANOVA do also hypothesis testing like your ANOVA. What is the hypothesis here what are the hypothesis in ANOVA.

(Refer Slide Time: 14:04)



You said $\mu_1 = \mu_2 = \dots = \mu_L$ that is your H_0 and your H_1 is $\mu_l \neq \mu_m$. For at least one pair of μ_l, μ_m one pair of μ means either μ_l or μ_m $l = 1, 2, \dots, L, m = 1, 2, \dots, L$ and $l \neq m$. In MANOVA case the same hypothesis is same, that we are saying $H_0: \mu_1 = \mu_2 = \dots = \mu_L$. Please, keep in mind by saying this we are saying like this μ_{11}, μ_{12} like this $\mu_{1p} = \mu_{21}, \mu_{22}, \mu_{2p}$ equal to finally, $\mu_{L1}, \mu_{L2}, \mu_{Lp}$ that is the difference in ANOVA case, it is scalar quantity MANOVA case it is the vector quantity. Your alternate hypothesis that $\mu_l \neq \mu_m$ by saying this you are saying that $\mu_{l1}, \mu_{l2}, \mu_{lp}$ this $\neq \mu_{m1}, \mu_{m2}$ for at least one pair of (l, m) .

Okay so, like ANOVA we will also partition the general observation. So, what is my general observation here?

(Refer Slide Time: 16:48)

$$X_{il} = \mu + (\mu_i - \mu) + (X_{il} - \mu_i)$$

$$= \mu + \gamma_i + \epsilon_{il}$$

Labels for the equation above:

- μ : grand mean
- γ_i : population effect
- ϵ_{il} : Random error

$$\gamma_i = \begin{bmatrix} \gamma_{i1} \\ \gamma_{i2} \\ \vdots \\ \gamma_{ip} \end{bmatrix}_{p \times 1}$$

$$X_{il} = \mu + \gamma_i + \epsilon_{il}$$

Labels for the vector equation:

- X_{il} : general observation vector
- μ : grand mean vector
- γ_i : population effect vector
- ϵ_{il} : Random error vector

Matrix representation of the partitioning:

$$\begin{bmatrix} X_{i1} \\ X_{i2} \\ \vdots \\ X_{ip} \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix} + \begin{bmatrix} \mu_{i1} - \mu_1 \\ \mu_{i2} - \mu_2 \\ \vdots \\ \mu_{ip} - \mu_p \end{bmatrix} + \begin{bmatrix} X_{i1} - \mu_{i1} \\ X_{i2} - \mu_{i2} \\ \vdots \\ X_{ip} - \mu_{ip} \end{bmatrix}$$

My general observation here is X_{il} . Let us see in ANOVA case, parallelly see MANOVA, case in ANOVA case X_{il} is partitioned like this, $\mu + \mu_i - \mu + X_{il} - \mu_i$, this is the way you partition and we say this is equal to $\mu + \tau_i + \epsilon_{il}$ μ is grand mean τ_i is the population effect. We are saying that no model is perfect, it cannot predict exactly same exact value for all the observations, there will be random errors. So, $\sum \epsilon_{il}$ is the random error part. Okay so, same partitioning possible in MANOVA, what we will write $X_{il} = \mu + \mu_i - \mu + X_{il} - \mu_i$.

So, it similar it is like this we have seen that X_{il} is $X_{i1} X_{i2} X_{ip}$. p variables are there, which is equivalent to $\mu_1 \mu_2$ like our case is μ_1 that is the grand mean case plus if we write like this think earlier we have written $\mu_{i1} - \mu_1 \mu_{i2} - \mu_2$ in the same manner you come μ , I think I this μ_1 . Now, $\mu_i - \mu$, so all cases I will be there and then every case what is, what we are doing. This is μ_1 this is μ_p not μ_1 this is μ_p , please write down this is μ_p . So, the μ_{ip} -s μ_p , that is why the problem comes, I have written I then plus same thing X_{il} all this you write $X_{i1} - \mu_{i1} X_{i2} - \mu_{i2}$, same manner you come $X_{ip} - \mu_{ip}$.

This is what is partitioning the general observation, to three components one is this one is the general observation to the left had side general observation vector and right hand side. This is

your grand mean vector then other one that $\mu_{i1}-\mu_1$. This one we are saying population effect vector population effect vector and last one is random error vector okay this partitioning for this one this partitioning we can write also in the same manner earlier we have written that this first one. This is X_{ij} fine then second one is μ that is also fine then third, it will be $\tau_1 + \epsilon_{ij}$.

So, this is my general observation, this my grand mean vector, this is the population effect vector, this is the random error effect okay so, if this is just we will just what I mean to say that let us write down the τ_1 case. This will be a vector $\tau_{11} \tau_{12}$ like this τ_{1p} , $p \times 1$ vector. So, when we say l^{th} population effect that is related to all the variables considered here we are considering p variables. So, τ_1 for p variables okay so, if we frame our hypothesis like this.

(Refer Slide Time: 23:02)

Handwritten mathematical derivation on a blue background:

$$\left[\begin{array}{l} H_0: \mu_1 = \mu_2 = \dots = \mu_L \\ H_1: \mu_l \neq \mu_m \end{array} \right. \left. \begin{array}{l} \leftarrow \\ \text{True} \end{array} \right.$$

$$\gamma_l = \mu_l - \mu \quad \mu = \frac{n_1 \mu_1 + n_2 \mu_2 + \dots + n_L \mu_L}{n_1 + n_2 + \dots + n_L}$$

$$\mu_l = \mu \text{ if } H_0 \text{ is true} = \frac{(n_1 + n_2 + \dots + n_L) \cdot \mu}{n_1 + n_2 + \dots + n_L}$$

$$\therefore \gamma_l = \mu_l - \mu = 0.$$

$$\left[\begin{array}{l} H_0: \gamma_l = 0, \quad l = 1, 2, \dots, L \\ H_1: \gamma_l \neq 0, \quad \text{for at least one } l. \end{array} \right.$$

That $H_0 \mu_1 = \mu_2$ equal to μ_1 and $H_1 \mu_1 \neq \mu_m$ for one pair of l_m . Then using this τ_1 concept, you can find out the null hypothesis also, what will be the null hypothesis also, what is τ_1 . τ_1 is $\mu_1 - \mu$. Now if H_0 is true if this one is true means all means are equal, then what will be the grand mean? Grand mean will be $n_1 \mu_1 + n_2 \mu_2 + \dots + n_L \mu_L / n_1 + n_2 + \dots + n_L$. If all means are equal then what will happen we can write all $\mu_1 = \mu_2$. So, it will $n_1 + n_2 + \dots + n_L / n_1 + n_2 + \dots + n_L$ into μ because this will be

cancelled out. So, what we mean then we want to say that every μ will be equal to the grand μ getting me.

So, then what we can say that $\mu_l = \mu$ if H_0 is true, which indicates $\tau_l = \mu_l - \mu = 0$. That means we can create null hypothesis like this $\tau_l = 0, l = 1, 2, L$ and your alternate hypothesis will be $\tau_l \neq 0$ for at least one l . So, if you test one hypothesis in terms of μ and other hypothesis in terms of τ_l you are actually doing the same thing. Okay so, in MANOVA.

(Refer Slide Time: 25:46)

Conceptual model: parameters

$$X_{il} = \mu + \tau_l + \epsilon_{il}$$

$$\tau_l = \mu_l - \mu \quad \epsilon_{il} = X_{il} - \mu_l$$

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

$$\text{and } \sum_{l=1}^L n_l \tau_l = 0, \text{ for unequal sample size}$$



We will do this in the same manner like ANOVA partitioning. Now, we partitioned the observation, now we will see that how to partition the this one your variability part, but what are the parameters you are estimating in MANOVA your parameters will be this τ_l as well μ_l . Also, you have to estimate μ you have to estimate and you have to estimate also the error terms and another issue here is if you go for unequal sample size, then the weighted effect of the population that sum will become zero. If you go for equal size that is again, the sum of the effects of the populations will become zero.

I think you can prove the second one also first one, why it is zero what I mean to say, we are saying that we are saying that.

(Refer Slide Time: 27:00)

$$\sum_{k=1}^L n_k \tau_k = 0, \quad \tau_k = \mu_k - \mu$$

$$n_k \tau_k = \sum_{k=1}^L n_k (\mu_k - \mu)$$

$$= \sum_{k=1}^L n_k \mu_k - \sum_{k=1}^L \mu$$

Some total of $n_1 \tau_1$ $l = 1$ to L this is zero you see, what is τ_1 is $\mu_1 - \mu$ So, if you write down here sum total of $l = 1$ to L n_1 into $\mu_1 - \mu$ the LH that is the left hand side, what you will get here this one $l = 1$ to L n_1 and $\mu_1 - l = 1$ to $L \times \mu$. So, you have already seen that.

(Refer Slide Time: 27:55)

Handwritten notes on a whiteboard:

$H_0: \mu_1 = \mu_2 = \dots = \mu_k$
 $H_1: \mu_l \neq \mu_m$

$\gamma_l = \mu_l - \mu$ $\mu = \frac{n_1 \mu_1 + n_2 \mu_2 + \dots + n_k \mu_k}{n_1 + n_2 + \dots + n_k}$

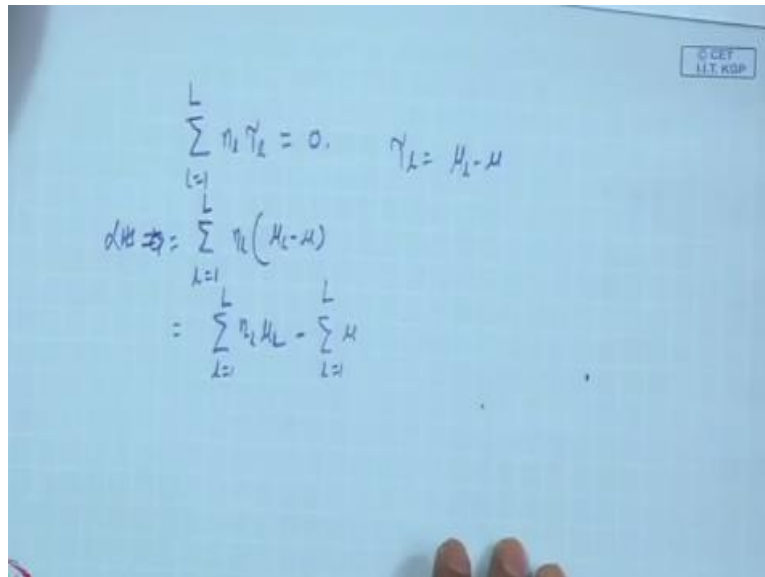
$\mu_l = \mu$ if H_0 is true

$\therefore \gamma_l = \mu_l - \mu = 0$

$H_0: \gamma_l = 0, l = 1, 2, \dots$
 $H_1: \gamma_l \neq 0, \text{ for at least one } l$

$\mu = n_1 \mu_1 + n_2 \mu_2 + n_1 \mu_1$ by this. So, that means the sum of this $n_1 \mu_1$ can be written like this.

(Refer Slide Time: 28:11)



The image shows a whiteboard with handwritten mathematical equations. In the top right corner, there is a small rectangular box containing the text "CCEET IIT KGP". The main content consists of the following equations:

$$\sum_{k=1}^L n_k \gamma_k = 0, \quad \gamma_k = \mu_k - \mu$$
$$\text{LHS} \Rightarrow \sum_{k=1}^L n_k (\mu_k - \mu)$$
$$= \sum_{k=1}^L n_k \mu_k - \sum_{k=1}^L n_k \mu$$

Can be written as this one can be written like this.

(Refer Slide Time: 28:18)

$$\left[\begin{array}{l} H_0: \mu_1 = \mu_2 = \dots = \mu_k \\ H_1: \mu_i \neq \mu_m \end{array} \right. \leftarrow \text{True}$$

$$\gamma_k = \mu_k - \mu$$

$$\mu_k = \mu \text{ if } H_0 \text{ is true}$$

$$\therefore \gamma_k = \mu_k - \mu = 0.$$

$$\left[\begin{array}{l} H_0: \gamma_k = 0, \text{ for } k=1, 2, \dots, k \\ H_1: \gamma_k \neq 0, \text{ for } k=1, 2, \dots, k \end{array} \right.$$

$$\mu = \frac{n_1 \mu_1 + n_2 \mu_2 + \dots + n_k \mu_k}{n_1 + n_2 + \dots + n_k}$$

$$= \frac{(n_1 + n_2 + \dots + n_k) \cdot \mu}{n_1 + n_2 + \dots + n_k}$$

This into this.

(Refer Slide Time: 28:24)

$$\begin{aligned} \sum_{k=1}^L n_k \gamma_k &= 0, \quad \gamma_k = \mu_k - \mu \\ \Rightarrow \sum_{k=1}^L n_k (\mu_k - \mu) & \\ &= \sum_{k=1}^L n_k \mu_k - \sum_{k=1}^L n_k \mu \\ &= \sum_{k=1}^L n_k \mu_k - (n_1 + n_2 + \dots + n_L) \mu \end{aligned}$$

Okay and then that will be cancelled out and you will become it will be here if it is n_1 is there you have not written this, so it is okay so, what we mean to say that this quantity is $n_1 + n_2 + n_1 \times \mu$, yes or no.

(Refer Slide Time: 28:54)

$$\begin{cases} H_0: \mu_1 = \mu_2 = \dots = \mu_k \\ H_1: \mu_1 \neq \mu_2 \neq \dots \neq \mu_k \end{cases}$$

$$T_k = \mu_k - \mu$$

$$\mu_k = \mu \text{ if } H_0 \text{ is true}$$

$$\therefore T_k = \mu_k - \mu = 0$$

$$\begin{cases} H_0: T_k = 0 \\ H_1: T_k \neq 0 \end{cases}$$

$$\mu = \frac{n_1 \mu_1 + n_2 \mu_2 + \dots + n_k \mu_k}{n_1 + n_2 + \dots + n_k} = \frac{(n_1 + n_2 + \dots + n_k) \cdot \mu}{n_1 + n_2 + \dots + n_k}$$

And I have told you this, that $n_1 + n_2 + n_1 \mu = n_1 \mu_1 + n_2 \mu_2 + n_1 \mu_1$.

(Refer Slide Time: 29:07)


The image shows a handwritten derivation on a blue background. At the top right, there is a small logo for 'CET IIT KGP'. The derivation starts with the equation $\sum_{k=1}^L n_k \gamma_k = 0$, where $\gamma_k = \mu_k - \mu$. This is then expanded to $\sum_{k=1}^L n_k (\mu_k - \mu)$. The next step is to separate the sum into $\sum_{k=1}^L n_k \mu_k - \sum_{k=1}^L n_k \mu$. The first term is written as $(n_1 \mu_1 + n_2 \mu_2 + \dots + n_L \mu_L)$ and the second term as $(n_1 + n_2 + \dots + n_L) \mu$. The difference is shown to be zero. Below this, the formula for the grand mean $\mu = \frac{n_1 \mu_1 + n_2 \mu_2 + \dots + n_L \mu_L}{n_1 + n_2 + \dots + n_L}$ is written, and it is noted that $(n_1 \mu_1 + n_2 \mu_2 + \dots + n_L \mu_L) = (n_1 + n_2 + \dots + n_L) \mu = 0$.

So, this again this none $n_1 \mu_1 + n_2 \mu_2 + n_1 \mu_1$ is this one $n_1 \mu_1 + n_2 \mu_2$. This is this part - $n_1 + n_2 + n_1 \mu$, this quantity equal to this quantity problem. You have to understand that the grand mean is mean of the means weighted case here, $n_1 \mu_1 + n_2 \mu_2 + n_1 \mu_1$ by total frequency and this one is μ . So, you can write from here $n_1 \mu_1 + n_2 \mu_2 + n_1 \mu_1 = n_1 + n_2 + n_1 \times \mu$. Then if you make minus this will become zero, this is the case.

(Refer Slide Time: 30:30)

Assumptions

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



Dr J Mishra, IEM, IIT Kharagpur

8

Okay now, we will see the assumptions what are the assumptions population covariances are equal, errors are normally distributed, errors are iid. We will see first population covariances are equal, how to test it.

(Refer Slide Time: 30:49)

Test of equality of population covariances: Box M test

Hypothesis $H_0 : \Sigma_1 = \Sigma_2 = \dots = \Sigma_L$
 $H_1 : \Sigma_\ell \neq \Sigma_m$, for at least pair of (ℓ, m) .

Statistic $D = (1 - u)M$

$$M = -2 \ln \left[\prod_{\ell=1}^L \left(\frac{|S_\ell|}{|S_{pooled}|} \right)^{(n_\ell - 1)/2} \right] = \left[\sum_{\ell} (n_\ell - 1) \ln |S_{pooled}| \right] - \sum_{\ell} [(n_\ell - 1) \ln |S_\ell|]$$

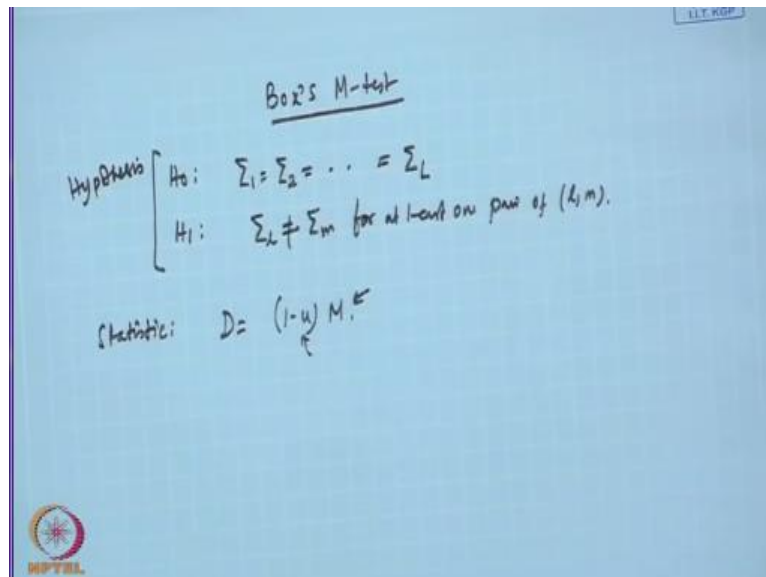
$$u = \left[\sum_{\ell} \frac{1}{(n_\ell - 1)} - \frac{1}{\sum_{\ell} (n_\ell - 1)} \right] \left[\frac{2p^2 + 3p - 1}{6(p+1)(L-1)} \right]$$

Decision Reject H_0 when $D > \chi_{\alpha, \nu}$. $\nu = \frac{1}{2} p(p+1)(L-1)$



We will be using box M test.

(Refer Slide Time: 31:04)



Box's M-test

Hypothesis $\left\{ \begin{array}{l} H_0: \Sigma_1 = \Sigma_2 = \dots = \Sigma_L \\ H_1: \Sigma_l \neq \Sigma_m \text{ for at least one pair of } (l,m). \end{array} \right.$

Statistic: $D = (1-u) M.$

Here we will create hypothesis H_0 that population covariance are equal and alternative hypothesis will be $\Sigma_l \neq \Sigma_m$, for at least one pair of l,m . Okay now, we create one statistic, so this is our hypothesis then creates the statistic our statistic we are creating suppose $D = 1 - u$ into M so, you require to know what is M and what is your u correct now, let us see the slide where I have written these things.

(Refer Slide Time: 32:22)

Test of equality of population covariances: Box M test

Hypothesis $H_0 : \Sigma_1 = \Sigma_2 = \dots = \Sigma_L$
 $H_1 : \Sigma_\ell \neq \Sigma_m$, for at least pair of (ℓ, m) .

Statistic $D = (1 - u)M$

$$M = -2 \ln \left[\prod_{l=1}^L \left(\frac{|S_l|}{|S_{pooled}|} \right)^{(n_l - 1)/2} \right] = \left[\sum_l (n_l - 1) \ln |S_{pooled}| \right] - \sum_l [(n_l - 1) \ln |S_l|]$$

$$u = \left[\sum_l \frac{1}{(n_l - 1)} - \frac{1}{\sum_l (n_l - 1)} \right] \left[\frac{2p^2 + 3p - 1}{6(p + 1)(L - 1)} \right]$$

Decision Reject H_0 when $D > \chi_{\alpha, \nu}$, $\nu = \frac{1}{2} p(p + 1)(L - 1)$



In slide you see that M is $-2 \log$ $l=1$ to L , L . That is the multiplication into that determinant of S_l by S_{pooled} to the power $n_l - 1 / 2$. This is what is actually the this ratio S_l / S_{pooled} all of you know S_{pooled} , what is S_{pooled} how to come to S_{pooled} .

(Refer Slide Time: 33:06)

Box's M-test

Hypothesis $\left\{ \begin{array}{l} H_0: \Sigma_1 = \Sigma_2 = \dots = \Sigma_L \\ H_1: \Sigma_L \neq \Sigma_m \text{ for at least one pair of } (L, m). \end{array} \right.$

Statistic: $D = (1-u) M$ ← $S_{\text{pooled}} = \frac{(n_1-1)S_1 + (n_2-1)S_2 + \dots + (n_L-1)S_L}{(n_1+n_2+\dots+n_L)-L}$

Your $n_1 - 1 S_1 + n_2 - 1 S_2 + n_1 - 1 S_1$ / what? $n_1 + n_2 + n_1$ - correct so, that you have see earlier also in two variable, sorry two population univariate case you have seen that $n_1 - 1 \times S_1 + n_2 - 1 \times n_1 + n_2 - 2$ just check, this is the case. Now, again you see the formula.

(Refer Slide Time: 33:59)

Test of equality of population covariances: Box M test

Hypothesis $H_0 : \Sigma_1 = \Sigma_2 = \dots = \Sigma_L$
 $H_1 : \Sigma_\ell \neq \Sigma_m$, for at least pair of (ℓ, m) .

Statistic $D = (1 - u)M$

$$M = -2 \ln \left[\prod_{\ell=1}^L \left(\frac{|S_\ell|}{|S_{pooled}|} \right)^{(n_\ell - 1)/2} \right] = \left[\sum_{\ell} (n_\ell - 1) \ln |S_{pooled}| \right] - \sum_{\ell} [(n_\ell - 1) \ln |S_\ell|]$$

$$u = \left[\sum_{\ell} \frac{1}{(n_\ell - 1)} - \frac{1}{\sum_{\ell} (n_\ell - 1)} \right] \left[\frac{2p^2 + 3p - 1}{6(p+1)(L-1)} \right]$$

Decision Reject H_0 when $D > \chi_{\alpha, \nu}$, $\nu = \frac{1}{2} p(p+1)(L-1)$



If your $\Sigma_1 = \Sigma_2 = \dots = \Sigma_L$, then your S_{pooled} will be $= S_1$ or S_L in general term. So, that means this determinant by this and determinant that will vary that will be ratio will be one and why have taken log. The log is taken to make it linear because, it is a multiplicative one to make it a linear one the log is taken here like this. So, if you have taken $-2 \log$ it is coming like this. This quantity is literalized like this $\sum_{\ell} (n_\ell - 1) \log |S_{pooled}| - \sum_{\ell} (n_\ell - 1) \log |S_\ell|$ this our m value. Okay so, this m value and what is the u value u is $\sum_{\ell} \frac{1}{(n_\ell - 1)} - \frac{1}{\sum_{\ell} (n_\ell - 1)}$ by that this sum into $\frac{2p^2 + 3p - 1}{6(p+1)(L-1)}$ into $1 - 1$.

Okay this the development by box. So, if you then you put m and u in D, now D follows χ^2 distribution with ν degrees of freedom, where ν is $\frac{1}{2} p \times p + 1 \times 1 - 1$ okay so, you have to remember this, what is your ν value.

(Refer Slide Time: 35:54)


Box's M-test

Hypothesis $\left\{ \begin{array}{l} H_0: \Sigma_1 = \Sigma_2 = \dots = \Sigma_L \\ H_1: \Sigma_L \neq \Sigma_m \text{ for at least one pair of } (L, m). \end{array} \right.$

Statistic: $D = (1-u) M_1^L$ $S_{\text{pooled}} = \frac{(n_1-1)S_1 + (n_2-1)S_2 + \dots + (n_L-1)S_L}{(n_1+n_2+\dots+n_L)-L}$

degrees of freedom for D: $v = \frac{1}{2} p(p+1)(L-1)$

$D > \chi^2_{\alpha, v} \leftarrow \text{Reject } H_0$



$v = \frac{1}{2} p \times p + 1 \times 1 - 1$, that is the degrees of freedom for D. So you if your D \geq equal to χ^2_{α} and v, then you reject H_0 population variances are not equal. We have calculated this for this data set.

(Refer Slide Time: 36:34)

Box M test


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

S ₁		S ₂		S ₃	
1.51	0.11	1.43	0.52	0.93	-0.11
0.11	1.17	0.52	0.68	-0.11	0.72

Spooled		M	1.04
1.29	0.17	U	0.11
0.17	0.86	D	0.93

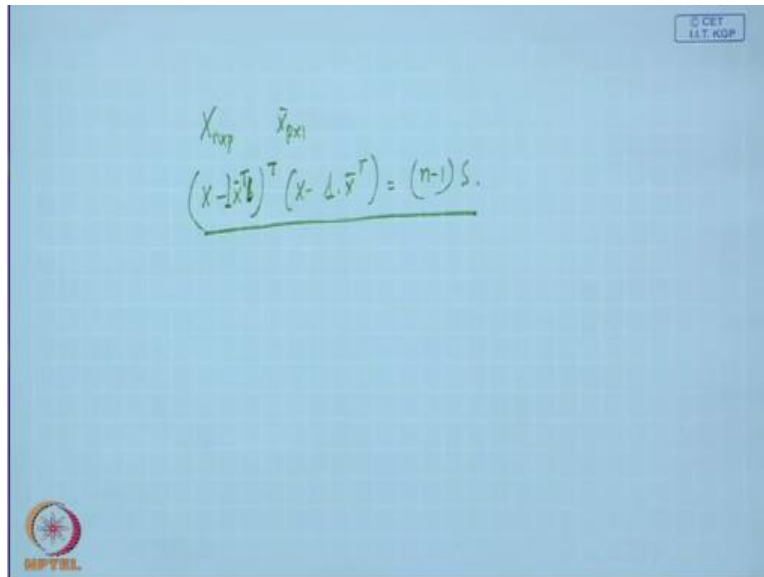
dof	6	Decision
chi-sq(6, 0.05)	12.59	Accept H ₀

$S_0, \Sigma_1 = \Sigma_2 = \Sigma_3 = \Sigma$


Dr J Mallick, IEM, IIT Kharagpur
10

Are you not comfortable now, to compute the covariance matrix for a given data set for process A the covariance matrix is S_1 , for process B it is S_2 , for process C it is S_3 you will be using can you recall the covariance matrix formula, what you have used.

(Refer Slide Time: 37:05)



The image shows a handwritten mathematical formula on a light blue grid background. The formula is $(X - \bar{x}\bar{x}^T)^T (X - \bar{x}\bar{x}^T) = (n-1)S$. Above the first term, the dimensions are noted as $X_{n \times p}$ and $\bar{x}_{p \times 1}$. The entire equation is underlined. In the top right corner, there is a small box containing the text "© CET I.I.T. KGP". In the bottom left corner, there is a circular logo with a star and the text "NPTEL" below it.

$$\begin{matrix} X_{n \times p} & \bar{x}_{p \times 1} \\ \underline{(X - \bar{x}\bar{x}^T)^T (X - \bar{x}\bar{x}^T) = (n-1)S.} \end{matrix}$$

If your X is $n \times p$ and your \bar{x} is $p \times 1$ you have created $x - \bar{x}$ you also multiplied by 1. So, to make it $n \times 1$, I think you have done like this 1, this transpose, then this one transpose $X - \bar{x}\bar{x}^T$ this will be $n - 1 \times S$ same formula we have used here.

(Refer Slide Time: 37:47)

Box M test


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

S ₁		S ₂		S ₃	
1.51	0.11	1.43	0.52	0.93	-0.11
0.11	1.17	0.52	0.68	-0.11	0.72

Spooled		M	1.04
1.29	0.17	U	0.11
0.17	0.86	D	0.93

dof	6	Decision
chi-sq(6, 0.05)	12.59	Accept H ₀

$S_0, \Sigma_1 = \Sigma_2 = \Sigma_3 = \Sigma$


Dr J Malvi, IEM, IIT Kharagpur
10

We found out S_1 for this, this row this column and this column this two column OD column for process A. ID column for process A is S_1 you are getting using same formula you get S_2 you get S_3 correct then what you require to know, you require to know S_{pooled} , S_{pooled} will be $S_1 + S_2 + S_3$ divided by because here $n_1 = n_2 = n_3$.

(Refer Slide Time: 38:25)

The image shows a handwritten derivation on a blue background. At the top, it states $X_{exp} \quad \bar{X}_{pki}$. Below this, the formula $(X - \bar{X})^T (X - \bar{X}) = (n-1)S$ is written and underlined. The next line shows $n_1 = n_2 = n_3 = n = 10$. The final line calculates the pooled variance: $S_{pooled} = \frac{(10-1)S_1 + (10-1)S_2 + (10-1)S_3}{10+10+10-3} = \frac{9}{27} (S_1 + S_2 + S_3) = \frac{1}{3} (S_1 + S_2 + S_3)$. In the bottom left corner, there is a small circular logo with the text 'MPTEL' below it.

So write down like this $n_1 = n_2 = n_3 = n = 10$. So, my S_{pooled} will be $n - 1$, that means $(10 - 1) S_1$, $(10 - 1) S_2$, $(10 - 1) S_3 / n_1$ is what, $10 + 10 + 10 - 3$. So, it is $9 / 27 (S_1 + S_2 + S_3)$, so $1 / 3 (S_1 + S_2 + S_3)$. Now, you see any one of the value, suppose I want to know 1.29, here.

(Refer Slide Time: 39:14)

Box M test


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

S ₁		S ₂		S ₃	
1.51	0.11	1.43	0.52	0.93	-0.11
0.11	1.17	0.52	0.68	-0.11	0.72

Spooled		M	1.04
1.29	0.17	u	0.11
0.17	0.86	D	0.93

dof	6	Decision
chi-sq(6, 0.05)	12.59	Accept H ₀

$S_0, \Sigma_1 = \Sigma_2 = \Sigma_3 = \Sigma$


Dr J Malvi, IEM, IIT Kharagpur
10

How 1.29 is coming 1.29 the corresponding values in S_1 is 1.51 in S_2 is 1.43 in S_3 is 0.93. So, you sum $1.51 + 1.43 + 0.93 / 3$ will give you this value, that you have calculated earlier also. Then using the formula for M that big formula you have calculated M value that is 1.04, $u = 0.11$ then $D = 1 - u \times M$, that is 0.93. Now what is your degree of freedom in this case for D I say that degrees of freedom are your μ , μ is $\frac{1}{2} p (p + 1) l - 1$.

(Refer Slide Time: 40:10)

$$\begin{aligned} X_{n \times p} & \quad \bar{X}_{p \times 1} \\ (X - \frac{1}{n} \bar{X} \mathbf{1}^T)^T (X - \frac{1}{n} \bar{X} \mathbf{1}^T) &= (n-1) S. \\ n_1 = n_2 = n_3 = n &= 10. \\ S_{pooled} &= \frac{(10-1) s_1 + (10-1) s_2 + (10-1) s_3}{10+10+10-3} \\ &= \frac{9}{27} (s_1 + s_2 + s_3) = \frac{1}{3} (s_1 + s_2 + s_3) \\ Y = \frac{1}{2} p(p+1)(L-1) &= \frac{1}{2} \times 2 \times 3 \times (3-1) \\ &= 3 \times 2 = \mathbf{6} \end{aligned}$$

So, $\mu = \frac{1}{2} p(p+1)(L-1)$ so $\frac{1}{2} p = 2 \times 3 \times$ what $3 - 1$. So, how that mean $3 \times 2 = 6$ So your degree of freedom for D is 6.

(Refer Slide Time: 40:28)

Box M test


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

S1		S2		S3	
1.51	0.11	1.43	0.52	0.93	-0.11
0.11	1.17	0.52	0.68	-0.11	0.72

Spooled		M	1.04
1.29	0.17	U	0.11
0.17	0.86	D	0.93

dof	6	Decision
chi-sq(6, 0.05)	12.59	Accept Ho

$S_0, \Sigma_1 = \Sigma_2 = \Sigma_3 = \Sigma$



Dr J Malvi, IEM, IIT Kharagpur

10

Now χ^2 6 with $\alpha 0.05$ that value is 12.59 you will be getting it from χ^2 table. Then you compare computed D value versus χ^2 tabulated value. Now, D value is 0.93, which is much less than 5., 12.59. So, we can say we are fail to reject null hypothesis. We are accepting null hypothesis that means the population covariances are equal. So, we have we have seen the equality of population covariances are satisfied. Okay if this is satisfied, then we will go for MANOVA.

(Refer Slide Time: 41:21)

Decomposition of total sum of squares

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

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

$$SSCP_B = \sum_{\ell=1}^L n_\ell (\bar{x}_\ell - \bar{x})(\bar{x}_\ell - \bar{x})^T$$

$$SSCP_E = (n_1 - 1)S_1 + (n_2 - 1)S_2 + \dots + (n_L - 1)S_L$$

$$SSCP_T = SSCP_B + SSCP_E$$

$$N-1 = L-1 + N-L \quad N = \sum_{\ell=1}^L n_\ell$$



So, now we will decompose it so this decomposition is simple, again it is not that tough, what the slide looks like is very difficult one, but it is not like this what we have seen in ANOVA.

(Refer Slide Time: 41:38)

ANOVA

$$X_{ij} = \mu + (\mu_1 - \mu) + (x_{ij} - \mu_1)$$

$$\hat{\mu} = \bar{x} \quad \hat{\mu}_1 = \bar{x}_1$$

$$x_{ij} = \bar{x} + (\bar{x}_1 - \bar{x}) + (x_{ij} - \bar{x}_1)$$

$$X_{ij} = \mu + (\mu_1 - \mu) + (x_{ij} - \mu_1)$$

$$x_{ij} = \bar{x} + (\bar{x}_1 - \bar{x}) + (x_{ij} - \bar{x}_1)$$

We say any observation when you collected X_{ij} that one is partitioned into, now again let me see from the population point of view I say $X_{ij} = \mu + \mu_1 - \mu + X_{ij} - \mu_1$ correct visible now, what is the estimate of μ that is \bar{x} what is the estimate of μ_1 that you have seen in MANOVA. That is \bar{x}_1 μ_1 , that is mean of the 1 population estimate is sample mean. So now, we are partitioning the sample observation X_{ij} which can be written like this $\bar{x} + \bar{x}_1 - \bar{x} + x_{ij} - \bar{x}_1$. That we have seen earlier same thing possible here in MANOVA. This is from ANOVA you have done. Now, from MANOVA you do MANOVA also we have seen that this vector is μ vector + $\mu_1 - \mu$ vector + $X_{ij} - \mu_1$.

That is the formulation and the estimates also will be like this so, we are writing a vector X_{ij} which is \bar{x} that is the sample mean vector + $\bar{x}_1 - \bar{x} + X_{ij} - \bar{x}_1$ okay so, you can write so, you have seen this one earlier, but it is.

(Refer Slide Time: 43:51)

$$\begin{aligned}x_{il} &= \bar{x} + (\bar{x}_i - \bar{x}) + (x_{il} - \bar{x}_i) \\x_{il} &= \mu + (\mu_i - \mu) + (x_{il} - \mu_i) \\x_{il} &= \bar{x} + (\bar{x}_i - \bar{x}) + (x_{il} - \bar{x}_i).\end{aligned}$$

$\left[\begin{array}{c} \vdots \\ \vdots \\ \vdots \end{array} \right]_{p \times 1} = \left[\begin{array}{c} \vdots \\ \vdots \\ \vdots \end{array} \right]_{p \times 1} + \left[\begin{array}{c} \vdots \\ \vdots \\ \vdots \end{array} \right]_{p \times 1} + \left[\begin{array}{c} \vdots \\ \vdots \\ \vdots \end{array} \right]_{p \times 1}$

What will happen is this one is $p \times 1$ equal to this will also be a $p \times 1$ + this difference $p \times 1$ + this difference, okay this is general partitioning of the sample observation you do little more manipulation here.

(Refer Slide Time: 44:28)

$$x_{il} - \bar{x} = (\bar{x}_l - \bar{x}) + (x_{il} - \bar{x}_l)$$

$$(x_{il} - \bar{x})(x_{il} - \bar{x})^T = \left[(\bar{x}_l - \bar{x}) + (x_{il} - \bar{x}_l) \right] \cdot \left[(\bar{x}_l - \bar{x}) + (x_{il} - \bar{x}_l) \right]^T$$

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

$$\sum_{i=1}^{n_l} x_{il} = n_l \bar{x}_l \quad \sum_{i=1}^{n_l} \bar{x}_l = n_l \bar{x}_l$$

What you will do now, we will write like this $X_{il} - \bar{x} = \bar{x}_l - \bar{x} + X_{il} - \bar{x}_l$. If you take square, what will happen yes, transpose because, this is the vector form. So, you require to make like this $X_{il} - \bar{x} \times X_{il} - \bar{x}^T = \bar{x}_l - \bar{x} + X_{il} - \bar{x}_l \times \bar{x}_l - \bar{x}^T$ correct so, our $X_{il} - \bar{x}$ is a $p \times 1$ matrix transpose will be a $1 \times p$ matrix and the resultant will be $p \times p$ matrix, that is what we want also. Okay now, how many how many dimensions you have consider one is i another one is l and other one is j , $l = 1, 2 \dots i = 1, 2, \dots n$ or n_l an elevate unequal sample size we will consider here and $j = 1$ to p .

So, we will make sum over this dimension first is with i , so if I make $\sum_{i=1}^{n_l}$, then this quantity will become $X_{il} - \bar{x} \times X_{il} - \bar{x}^T$. This will be if you multiplied this into this, this into this like this. So, I am into this + I can write $i = 1$ to n_l $\bar{x}_l - \bar{x}$ into this one, $X_{il} - \bar{x}_l^T$. So, first one to second one here + sum total of $i = 1$ to n_l going to the second one $X_{il} - \bar{x}_l (\bar{x}_l - \bar{x})^T$ + sum total $i = 1$ to n_l $X_{il} - \bar{x}_l X_{il} - \bar{x}_l^T$. See this one, second one $\bar{x}_l - \bar{x}$ into this $X_{il} - \bar{x}_l$ and the third one $X_{il} - \bar{x}_l \bar{x}_l - \bar{x}$, so this value $\bar{x}_l - \bar{x}$ is independent of i . Similarly, here $\bar{x}_l - \bar{x}$ is independent of i . So, that mean $\sum_{i=1}^{n_l}$ will be affected here $X_{il} \bar{x}_l$ as well as here.

If anyone you take $i = 1$ to n_l \bar{x}_l is nothing but $n \bar{x}_l$ I'm repeating $i = 1$ to n_l \bar{x}_l is nothing but $n \bar{x}_l$ means. What I mean to say here I am saying $\sum_{i=1}^{n_l} \bar{x}_l = n \bar{x}_l$ Now, again $\sum_{i=1}^{n_l}$ of $i = 1$ to n_l \bar{x}_l if

this also $n_i \bar{x}_i$. So, this will become because this is independent of i . So, this quantity becomes 0, similarly this quantity with this becomes 0. So, the two middle terms will be deleted, because they are 0.

(Refer Slide Time: 49:49)

$$\sum_{k=1}^L \sum_{i=1}^{n_k} (x_{ik} - \bar{x})(x_{ik} - \bar{x})^T = \sum_{k=1}^L \sum_{i=1}^{n_k} (\bar{x}_k - \bar{x})(\bar{x}_k - \bar{x})^T + \sum_{k=1}^L \sum_{i=1}^{n_k} (x_{ik} - \bar{x}_k)(x_{ik} - \bar{x}_k)^T$$

$$= \sum_{k=1}^L n_k (\bar{x}_k - \bar{x})(\bar{x}_k - \bar{x})^T + \sum_{k=1}^L \sum_{i=1}^{n_k} (x_{ik} - \bar{x}_k)(x_{ik} - \bar{x}_k)^T$$

$$\begin{bmatrix} SSCP_T \\ \vdots \end{bmatrix}_{p \times p} = \begin{bmatrix} SSCP_B \\ \vdots \end{bmatrix}_{p \times p} + \begin{bmatrix} SSCP_E \\ \vdots \end{bmatrix}_{p \times p}$$

$$N-1 \quad L = L-1 \quad + \quad N-L$$

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

So, then resultant equation will be like this $i = 1$ to $n_l \bar{x}_{il} - \bar{x}^T = i = 1$ to $n_l \bar{x}_l - \bar{x} \bar{x}_l - \bar{x}^T +$ sum total $i = 1$ to $n_l x_{il} - \bar{x}_{il} \bar{x}_l \times \bar{x}_{il} - \bar{x}_l^T$ correct now, this quantity can be further written like this. See here there in no i^{th} term, so you can straight away write $n_l (\bar{x}_l - \bar{x}) \bar{x}_l - \bar{x}_l^T +$ this i^{th} term is available here, $n_l x_{il} - \bar{x}_{il}$ and $x_{il} - \bar{x}_l^T$. So, we have taken sum over i , now we take sum over l . So, $l = 1$ to L then here it will be here also it will $l = 1$ to L here it will be $l = 1$ to L . So, $l = 1 = L$ then here l equal to $1 = L$.

So, do we require the Σp again we do not require, because we are doing everything in the matrix domain and the vector quantity has taken care of the number of variables. So, we do not require further sum, so what is this quantity. Now, left hand side quantity this is if you consider x_{il} and \bar{x} as a scalar quantity. Then this one is a square quantity and this square quantity. From all the observations point of view what you have seen in ANOVA. So, that you have seen in ANOVA, this one is SST sum square total, but here it is a vector quantity. When you are multiplying this

vector with its transpose in such a manner, it is creating a matrix not a scalar creating a matrix of $p \times p$ dimension.

So, we will write this as this one will be something like this $p \times p$ here will be $1 \times p$ cross $p \times 1$ + this also will become another $p \times p$ correct so, diagonal elements will be the variance part off diagonal will be the covariance part variability and covariability. Where here it is getting here $i=1$ to n_i and $l=1$ to L okay so, this one is SSCP total, this SSCP what is this there between population mean vector to the grand mean vector. So, we will write that is between then this one is error SSCP error. So, the total covariance matrix okay it is not actually the covariance, covariance that will be divided by the degrees of freedom.


So, we can write that total sum of squares product matrix is divided into two sources of variability, one is the population other one is the errors. So, total sum square cross product is equal to that between sum square cross product plus error sum square cross product. This is the difference from ANOVA big difference from ANOVA. In ANOVA you will be getting scalar quantity everywhere. Okay then what will be the degrees of freedom for this one it is $N - 1 = L - 1 +$ difference $N - L$ same thing what you have done in ANOVA.

So, when N equal to what sum of $l = 1$ to L n_l that is all the observations together. So, in ANOVA we partition SST into SSB and SSE, in MANOVA we partition the sum square cross product matrix of the total to between population and error correct so, when you require to calculate.

(Refer Slide Time: 56:25)

$$\begin{aligned}
 \sum_{k=1}^L \sum_{i=1}^{n_k} (x_{ki} - \bar{x})(x_{ki} - \bar{x})^T &= \sum_{k=1}^L \sum_{i=1}^{n_k} (\bar{x}_k - \bar{x})(\bar{x}_k - \bar{x})^T + \sum_{k=1}^L \sum_{i=1}^{n_k} (x_{ki} - \bar{x}_k)(x_{ki} - \bar{x}_k)^T \\
 &= \sum_{k=1}^L n_k (\bar{x}_k - \bar{x})(\bar{x}_k - \bar{x})^T + \sum_{k=1}^L \sum_{i=1}^{n_k} (x_{ki} - \bar{x}_k)(x_{ki} - \bar{x}_k)^T
 \end{aligned}$$

$$\begin{bmatrix} \text{SSCP}_T \\ \vdots \end{bmatrix}_{p \times p} = \begin{bmatrix} \text{SSCP}_B \\ \vdots \end{bmatrix}_{p \times p} + \begin{bmatrix} \text{SSCP}_E \\ \vdots \end{bmatrix}_{p \times p}$$

$$\begin{matrix} N-1 & L \\ N = \sum_{k=1}^L n_k & = & L-1 & + & N-L \end{matrix}$$


SSCPT, SSCPB and SSCPE it is really difficult you see that in terms of matrix transpose then one sum by second sum like this. So, for computation point of view this one SSCPB is little easier than the other two. So, first you compute SSCPB using this formula absolutely.

(Refer Slide Time: 56:54)

✓ $SSCP_B = \sum_{k=1}^L n_k (\bar{x}_k - \bar{x})(\bar{x}_k - \bar{x})^T$

✓ $SSCP_E = (n_1 - 1) S_1 + (n_2 - 1) S_2 + \dots + (n_L - 1) S_L$

✓ $SSCP_T = SSCP_B + SSCP_E$

Decomposition of $SSCP$ matrix ✓

No problem $SSCP_B$ computation will be like this, $l = 1$ to L $n_l \bar{x}_l - \bar{x}$ and $\bar{x}_l - \bar{x}^T$ correct then for $SSCP_E$ there is a formula, which is $n_1 - 1 S_1 + n_2 - 1 S_2$ like this $n_1 - 1 S_L$, okay you have seen in pooled covariance case this was divided by degree of freedom, but it is not a covariance one it is basically $SSCP$ matrix. So that degrees of freedom is not divided, so it is S_1 is to S_L all you can compute very easily. So, $SSCP_E$ will be computed $SSCP_B$ will also be computed formula. Then you compute $SSCP_T$, that is $SSCP_B + SSCP_E$ this is the these are the steps, basically first you compute this, compute this, then compute this.

Okay so, this is what is our decomposition of covariance matrices decomposition of I can say instead of covariance matrix. Although, it is basically the same way covariance matrix will come ultimately, but it is sum square and cross product matrix you write, I am writing $SSCP$ matrix that is better, so $SSCP$ matrix total to this two quantity. So, I think today we will stop here and next class I will show you the MANOVA table then all the tests, how to go for hypothesis testing. Then comparison, pair-wise comparison and other things okay thank you very much.

NPTEL Video Recording Team

NPTEL Web Editing Team

**Technical Superintendents
Computer Technicians**

A IIT Kharagpur Production

www.nptel.iitm.ac.in

Copyrights Reserved