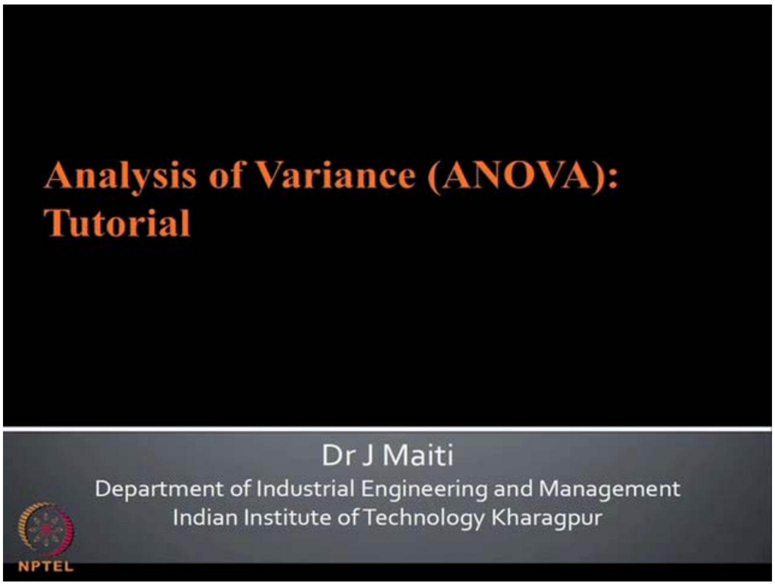
Applied Multivariate Statistical Modeling
Prof. J. Maiti
Department of Industrial Instrumentation and Management
Indian Institute of Technology, Kharagpur

Lecture - 18
Tutorial – ANOVA

(Refer Slide Time: 00:22)

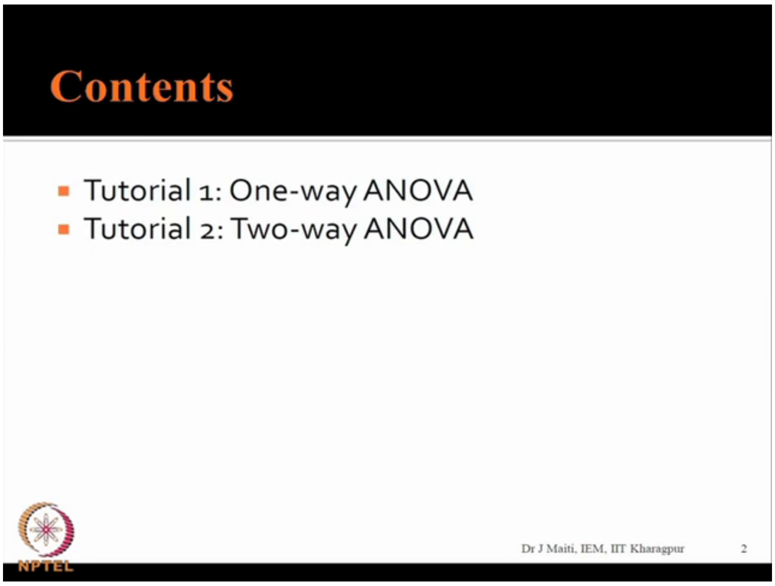


Analysis of Variance (ANOVA): Tutorial

Dr J Maiti
Department of Industrial Engineering and Management
Indian Institute of Technology Kharagpur

Good morning, today I will discuss. Tutorial on ANOVA we have two tutorials today.

(Refer Slide Time: 00:27)



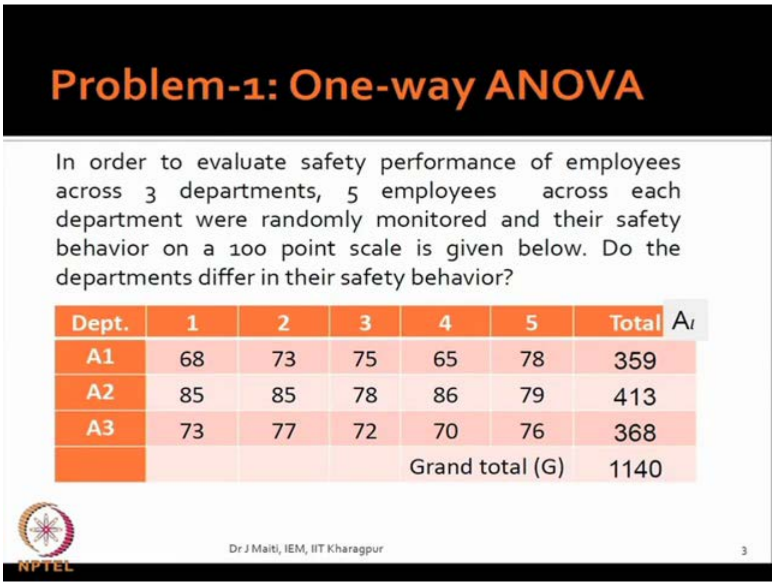
Contents

- Tutorial 1: One-way ANOVA
- Tutorial 2: Two-way ANOVA

Dr J Maiti, IEM, IIT Kharagpur 2

One is one tutorial one of one way ANOVA, and two way ANOVA.

(Refer Slide Time: 00:36)



Problem-1: One-way ANOVA

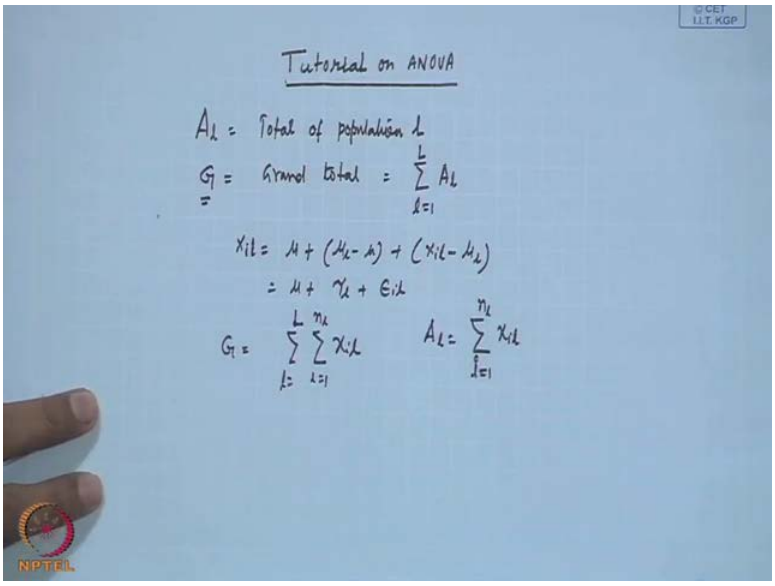
In order to evaluate safety performance of employees across 3 departments, 5 employees across each department were randomly monitored and their safety behavior on a 100 point scale is given below. Do the departments differ in their safety behavior?

Dept.	1	2	3	4	5	Total
A1	68	73	75	65	78	359
A2	85	85	78	86	79	413
A3	73	77	72	70	76	368
				Grand total (G)		1140

Dr J Maiti, IEM, IIT Kharagpur 3

Let us see the tutorial first one, one way ANOVA. The problem is in order to evaluate safety performance of employees across 3 departments, 5 employees across each department were randomly monitored. Their safety behavior on 100 point scale is given below do the departments differ in their safety behavior, this is our question. On the data point you see that there are three departments A 1, A 2, A 3; each departments we have collected five data points 1 to 5, and these data are 68, 73 like this. Now, we will see that it will some way or another with respect to this data set, if you make the total across A 1 the total score on safety behavior is 359 across A 2, it is 413 and across A 3 is 368 and the grand total is 1140.

(Refer Slide Time: 01:46)



Tutorial on ANOVA

A_l = Total of population l

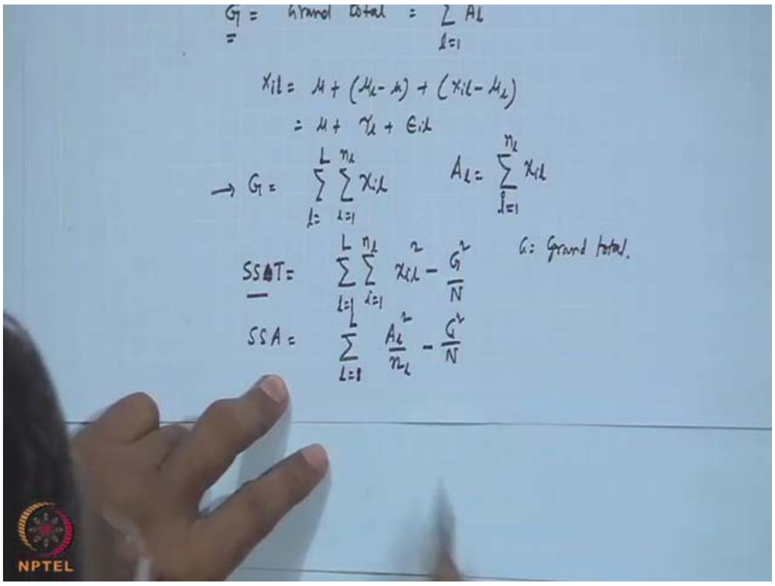
G = Grand total $= \sum_{l=1}^{L} A_l$

$$x_{il} = \mu + (\mu_l - \mu) + (x_{il} - \mu_l)$$
$$= \mu + \tau_l + \epsilon_{il}$$

$$G = \sum_{l=}^{L} \sum_{i=1}^{n_l} x_{il} \qquad A_l = \sum_{i=1}^{n_l} x_{il}$$

So, we are using adding little some different notation now, I am saying that A l is the total score or the total values total of population l, and we are saying G equal to the grand total, grand total this will be definitely l equal to 1 into L A l. So, what is our ANOVA model, here ANOVA model is x i l equal to mu plus mu l minus mu plus x i l minus mu l which is mu plus tau l plus epsilon i l this is our ANOVA model we have seen and we will now see the computation of al and G. So, if you what is what will be the grand total, grand total will be sum of I think I have already given you x i l i equal to 1 to n l, l equal to 1 to l and will be the grand mean and what will be the sum of this a l l equal to, so i equal to 1 to n l and x i l.

(Refer Slide Time: 03:31)



$$G = \text{Grand total} = \sum_{l=1} A_l$$

$$x_{il} = \mu + (\mu_l - \mu) + (x_{il} - \mu_l)$$
$$= \mu + \tau_l + \epsilon_{il}$$

$$\rightarrow G = \sum_{l=}^{L} \sum_{i=1}^{n_l} x_{il} \qquad A_l = \sum_{i=1}^{n_l} x_{il}$$

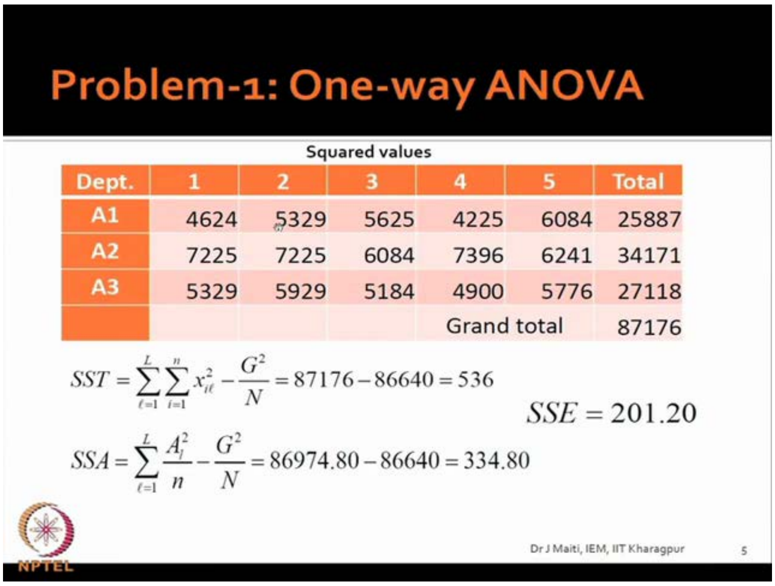
$$\underline{S}S\cancel{A}T = \sum_{l=1}^{L} \sum_{i=1}^{n_l} x_{il}^2 - \frac{G^2}{N} \qquad G: \text{Grand total.}$$

$$SSA = \sum_{l=1}^{L} \frac{A_l^2}{n_l} - \frac{G^2}{N}$$

NPTEL

Then we have seen what is the computation of S S A that is sum square total here we will use, here we will use little different notation computation for each of computation. That is this one you write down i equal to 1 to n l l equal to 1 to L x i l square minus we are saying that G square by n. We write down S S A equals to l equal to 1 to L i equal to 1 to n l x i l square minus G square by n where G is the grand total, grand total and this is our this is not S S A this is our S S T, this one is S S T sum square total. Now, sum square A means the population we are defining the population A this one is, now l equal to 1 to L a l square divided by n l minus G square by n.

(Refer Slide Time: 05:08)



Problem-1: One-way ANOVA

Squared values

Dept.	1	2	3	4	5	Total
A1	4624	5329	5625	4225	6084	25887
A2	7225	7225	6084	7396	6241	34171
A3	5329	5929	5184	4900	5776	27118
				Grand total		87176

$$SST = \sum_{\ell=1}^{L} \sum_{i=1}^{n} x_{i\ell}^2 - \frac{G^2}{N} = 87176 - 86640 = 536$$

$$SSE = 201.20$$

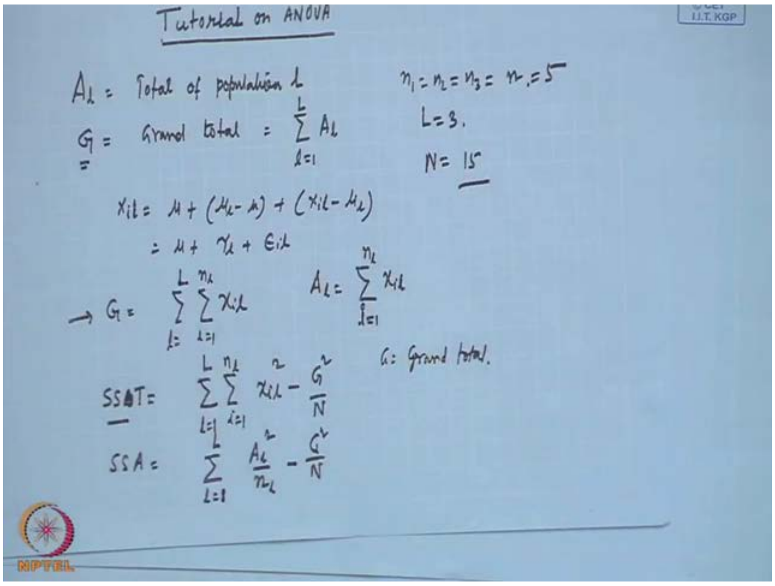
$$SSA = \sum_{\ell=1}^{L} \frac{A_l^2}{n} - \frac{G^2}{N} = 86974.80 - 86640 = 334.80$$

NPTEL

Dr J Maiti, IEM, IIT Kharagpur 5

So, this one if you see, here in this slide what we have written these are all the earlier values that these values 68, 73, 75 like this these values are square, these are squared values and this is the square the sum squared across A 1 across A 2 across A 3. Finally, this is the grand sum squared values, so S S T calculation formula is l equal to 1 to l i equal to 1 to n x i l square minus this and now this value is square sum square of all these values. So, it is 87,176 minus G square G already we have found out, here what is the value of G, G is 1140, so 1140 square by n what will be the value of n, n is l into n sum of n l into l.

(Refer Slide Time: 06:15)



Tutorial on ANOVA

A_l = Total of population l

G = Grand total $= \sum_{l=1}^{L} A_l$

$n_1 = n_2 = n_3 = n = 5$

$L = 3.$

$N = 15$

$x_{il} = \mu + (\mu_l - \mu) + (x_{il} - \mu_l)$

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

$\rightarrow G = \sum_{l=}^{L} \sum_{i=1}^{n_l} x_{il}$ $\quad A_l = \sum_{i=1}^{n_l} x_{il}$

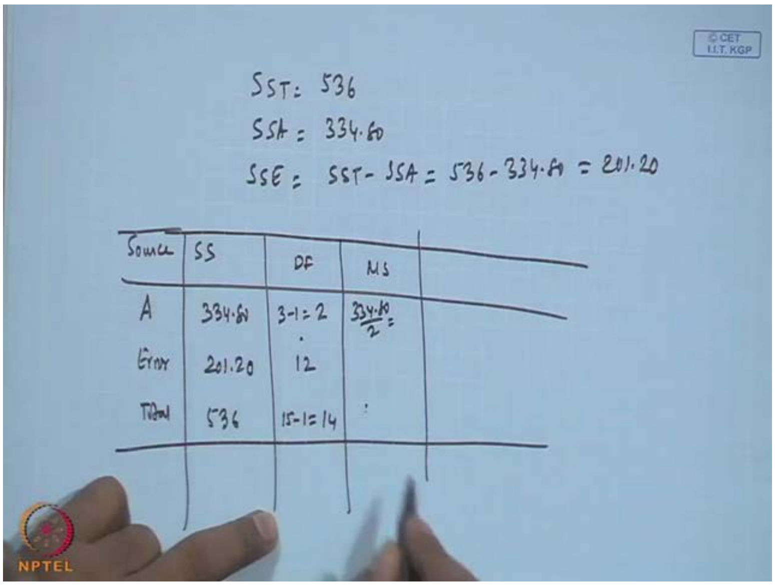
G: Grand total.

$SST = \sum_{l=1}^{L} \sum_{i=1}^{n_l} x_{il}^2 - \frac{G^2}{N}$

$SSA = \sum_{l=1}^{L} \frac{A_l^2}{n_l} - \frac{G^2}{N}$

So, here in this case our n 1 equal to n 2 equal to n 3 equal to n and l equal to 3 and n equal to also 5, so N will be 15, so that is why what is happening here G square by n 1140 square by 15 and this quantity is 536. Now, S S A will be this one l equal to L A, A l square by n you see what is this A l square A 1 square 25,887 A 2 square is 34,171 and third one A 3 square which is 27 and 118 this sum divided by n and then minus G square by N and the quantity will be 334.

(Refer Slide Time: 07:15)



SST: 536

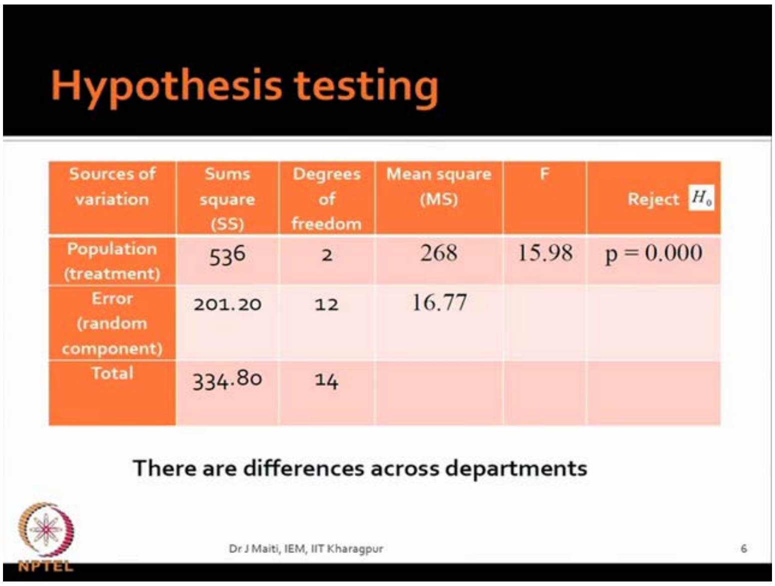
SSA = 334.80

SSE = SST - SSA = 536 - 334.80 = 201.20

Source	SS	DF	MS
A	334.80	3-1 = 2	$\frac{334.80}{2}$ =
Error	201.20	12	
Total	536	15-1 = 14	

So, then what we have got, now S S T equal to 536, S S A equal to 334.80 and S S E equal to S S T minus S S A which is 536 minus 334.80 which is 201.20. So, we have already seen we have a table for this sources of variation then your S S first is source, source is here population a then error then total then S S A is our 334.80 S S E is 201.20 and S S T is 536. What is our degrees of freedom, A 3 minus 1 equal to 2 and this will be 15 minus 1 equal to 14 then this will become 12 then what is our M S value 334.80 by 2, 334.80 by 2.

(Refer Slide Time: 08:36)



Hypothesis testing

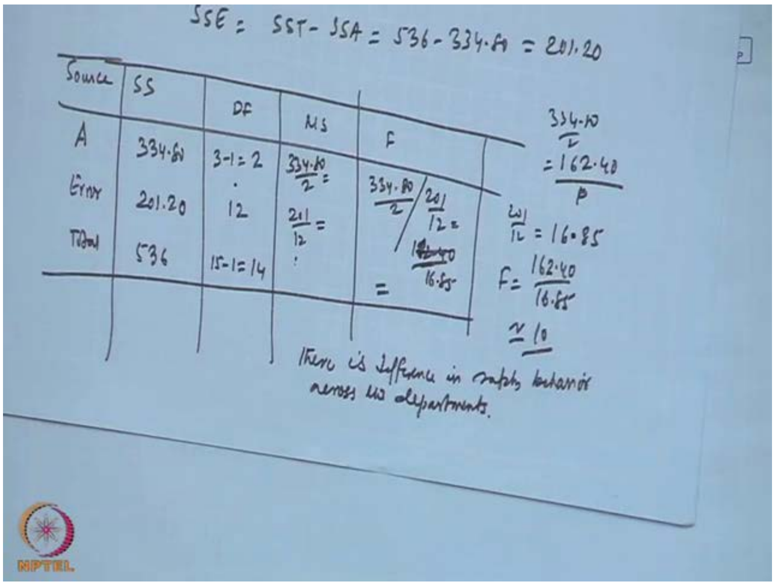
Sources of variation	Sums square (SS)	Degrees of freedom	Mean square (MS)	F	Reject H_0
Population (treatment)	536	2	268	15.98	p = 0.000
Error (random component)	201.20	12	16.77		
Total	334.80	14			

There are differences across departments

Dr J Maiti, IEM, IIT Kharagpur 6

What will be this value, 334.80 by 2, so this is S S T is 530, S S A is 334.80 by 2 you write down.

(Refer Slide Time: 08:58)



$SSE = SST - SSA = 536 - 334.80 = 201.20$

Source	SS	DF	MS	F
A	334.80	3-1=2	$\frac{334.80}{2} =$	$\frac{334.80}{2} \Big/ \frac{201}{12} = \frac{\cancel{162.40}}{16.85}$ =
Error	201.20	12	$\frac{201}{12} =$	
Total	536	15-1=14		

$\frac{334.80}{2} = 162.40$

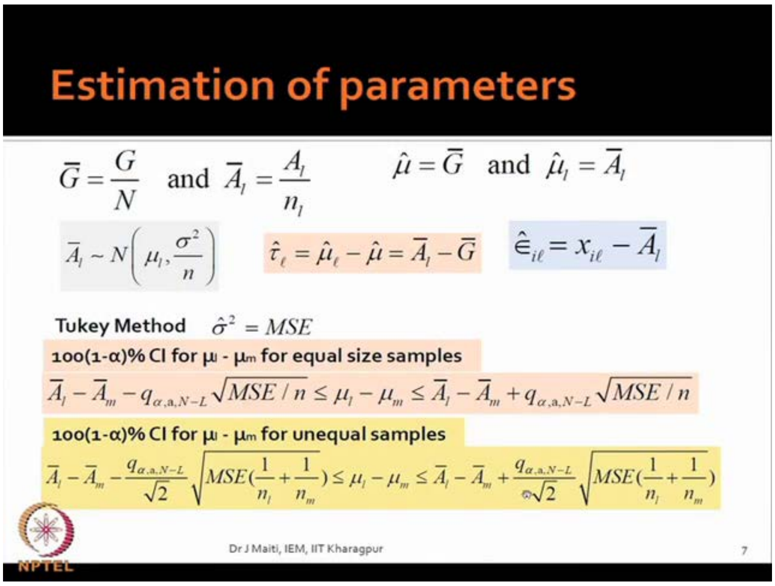
$\frac{201}{12} = 16.85$

$F = \frac{162.40}{16.85} \simeq 10$

There is difference in safety behavior across the departments.

Then 201 by 12 and then you write down that F statistics will be 334.80 by 2 divided by 201 by 12 some value will come that is already seen. So, this one, this total calculation if I made it will ultimately, it will be like this 334 by 2 means 334.80 by 2 means 162.40 divided by 201 by 12 that will be 1, then 881 then 6 into 85 something like this. So, the F will be 162.80 by 16.85 not coming, so F is F is 162.40 by 16.85 it will be almost equal to 10. It simply indicates that there is there is difference in safety behavior across the departments, so once that difference is established what is required to know we will go for estimation of parameters.

(Refer Slide Time: 11:04)



Estimation of parameters

$\overline{G} = \frac{G}{N}$ and $\overline{A}_l = \frac{A_l}{n_l}$ $\hat{\mu} = \overline{G}$ and $\hat{\mu}_l = \overline{A}_l$

$\overline{A}_l \sim N\left(\mu_l, \frac{\sigma^2}{n}\right)$ $\hat{\tau}_\ell = \hat{\mu}_\ell - \hat{\mu} = \overline{A}_l - \overline{G}$ $\hat{\epsilon}_{i\ell} = x_{i\ell} - \overline{A}_l$

Tukey Method $\hat{\sigma}^2 = MSE$

100(1-α)% CI for μ_l - μ_m for equal size samples

$$\overline{A}_l - \overline{A}_m - q_{\alpha,a,N-L}\sqrt{MSE/n} \le \mu_l - \mu_m \le \overline{A}_l - \overline{A}_m + q_{\alpha,a,N-L}\sqrt{MSE/n}$$

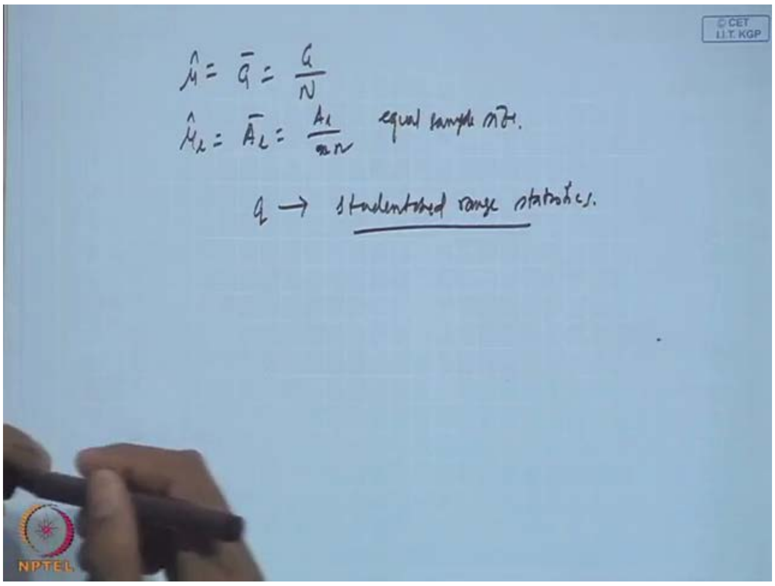
100(1-α)% CI for μ_l - μ_m for unequal samples

$$\overline{A}_l - \overline{A}_m - \frac{q_{\alpha,a,N-L}}{\sqrt{2}}\sqrt{MSE(\frac{1}{n_l}+\frac{1}{n_m})} \le \mu_l - \mu_m \le \overline{A}_l - \overline{A}_m + \frac{q_{\alpha,a,N-L}}{\sqrt{2}}\sqrt{MSE(\frac{1}{n_l}+\frac{1}{n_m})}$$

Dr J Maiti, IEM, IIT Kharagpur 7

Now, you see this slide this is what is estimation of parameter first one is we want to estimate the grand mean which is the G bar.

(Refer Slide Time: 11:17)



$\hat{\mu} = \overline{G} = \frac{G}{N}$

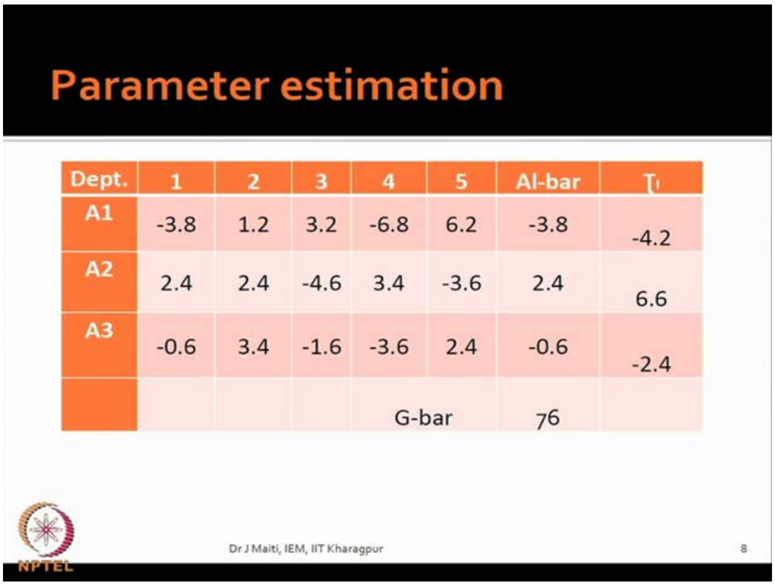
$\hat{\mu}_l = \overline{A}_l = \frac{A_l}{n}$ equal sample size.

$q \rightarrow$ studentised range statistics.

G bar grand mean estimated G bar that is G by N and your individual mean estimate is, here it will be your A l bar A l by I think it is A l by n because all cases. So, we have taken equals sample size and we are showing, here one method which is known as turkey method or turkey method, so this turkey method is used to see the difference in the mean values, the confidence interval as well as hypothesis testing. That is whether the difference is statistically significant or not and what is the interval confidence interval for this differences and this is simultaneous confidence interval.

Please keep in mind that, here one statistics is there q alpha A N minus l which is known as studentized range statistics, the q statistics here for turkey use is studentized range statistics. So, it is similar to other statistics like z and i square and other, but the values will be different, here alpha is the probability A and N minus l these are the other two parameters which will be used for seeing the table to see the studentized range statistics table turkey statistics is popularly used. Once you use, you can go for this interval estimation for equal sample sizes it will be like this and unequal sample sizes that one by N l plus 1 by N l this quantity will come into consideration.

(Refer Slide Time: 13:23)



Parameter estimation

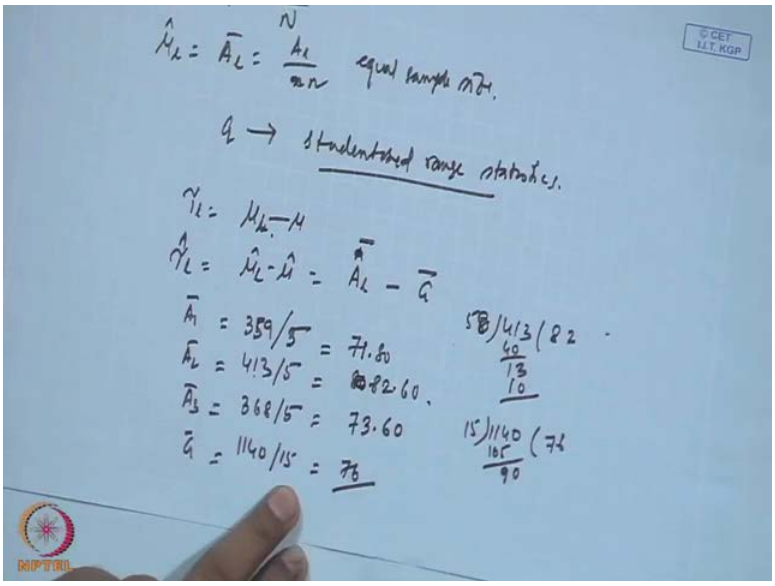
Dept.	1	2	3	4	5	Al-bar	τ_l
A1	-3.8	1.2	3.2	-6.8	6.2	-3.8	-4.2
A2	2.4	2.4	-4.6	3.4	-3.6	2.4	6.6
A3	-0.6	3.4	-1.6	-3.6	2.4	-0.6	-2.4
					G-bar	76	

Dr J Maiti, IEM, IIT Kharagpur

8

Now, let us see for the data points what we have what we have discussed, here our data point is our data point is these are the data points given. Now, with respect to this we want to see all the parameters, here tau l that is tau 1, tau 2, tau 3 what will be the tau L.

(Refer Slide Time: 13:54)



$\hat{\mu}_L = \bar{A}_L = \frac{A_L}{n}$ equal sample size.

$q \rightarrow$ studentized range statistics.

$\tau_L = \mu_L - \mu$

$\hat{\tau}_L = \hat{\mu}_L - \hat{\mu} = \bar{A}_L - \bar{G}$

$\bar{A}_1 = 359/5 = 71.80$

$\bar{A}_2 = 413/5 = 82.60.$

$\bar{A}_3 = 368/5 = 73.60$

$\bar{G} = 1140/15 = 76$

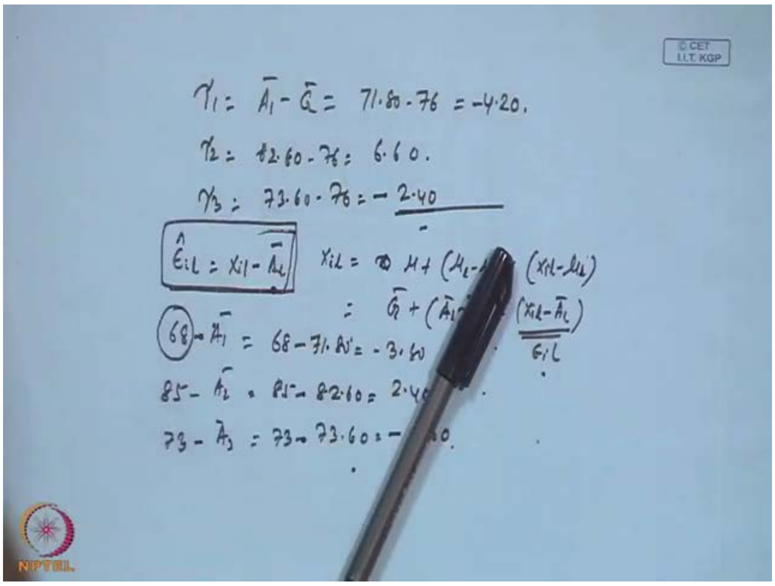
5)413(82
40
13
10

15)1140(76
105
90

Tau L is mu L minus mu, so if I say tau l estimate is mu l estimate minus mu estimate what is your tau l value, here A L bar minus G bar, so we know all those values we know what is your A L bar? This is A L 359. So, 359 then 413 then 368 and then the grand total is 1140, so here you divide it by 3, 5 you divide it by 5, here you divide it by 5, here you divide it by 5, here you divide it by 15.

So, then what will happen that A 1 bar, A 2 bar, A 3 bar and G bar equal to this, this, this what will be this value 359. So, it will be 71.80 second will be 80 point what will be this value 8, 4, 1, 3, 8, 5 this will be 8, 40, 1, 3 then 2, 10 that is 82, 82.60 and this one will be 5 into 7, 35, 5 into 3 then this 0 and this will be 1140, 15. So, 15 into I think it will 7, 105 then 90, 15 into 6, 76, now you come to this values here what will be your, this one that A 1 bar is given G is given the difference is tau 1.

(Refer Slide Time: 16:09)



$\tau_1 = \bar{A}_1 - \bar{G} = 71.80 - 76 = -4.20.$

$\tau_2 = 82.60 - 76 = 6.60.$

$\tau_3 = 73.60 - 76 = -2.40$

$\boxed{\hat{\epsilon}_{il} = x_{il} - \bar{A}_l}$ $x_{il} = \mu + (\mu_l -$ [illegible] $(x_{il} - \mu_l)$

$= \bar{G} + (\bar{A}_l$ [illegible] $\underbrace{(x_{il} - \bar{A}_l)}_{\epsilon_{il}}$

$68 - \bar{A}_1 = 68 - 71.80 = -3.80$

$85 - \bar{A}_2 = 85 - 82.60 = 2.4$[illegible]

$73 - \bar{A}_3 = 73 - 73.60 = -$[illegible]$0.$

© CET
I.I.T. KGP
NPTEL

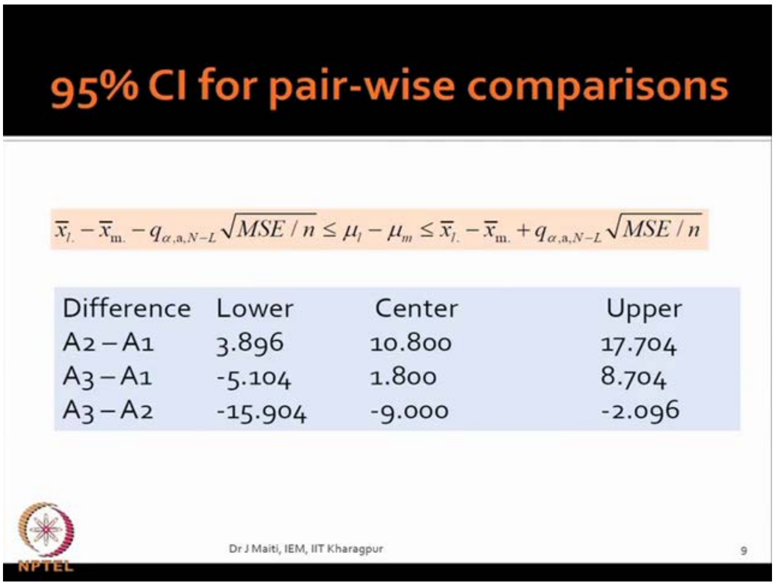
So, what will be my tau 1, tau 1 will be my A 1 bar minus G bar which is 71.80 minus 76 is it for minus 4.02 similarly tau 2 bar will be 82.60 minus 76, 6.60 and tau 3 bar will be 73.60 minus 76. So, minus 2.40 and sum total will be 0 then what you want, we want to calculate also the error terms what we have seen is that x i l is equal to your mu plus mu l minus mu plus x i l minus mu l what is nothing but G bar plus a l bar minus G bar plus x i l bar minus this is mu l. So, A l bar, so that means this one is 1 epsilon i l epsilon, i l is each of the values minus A l bar, now see you have already computed A l bar, A l bar is 80, 71.80 for A 1 for A 2 it is 82.6 and 3 is this the values are there.

So, you go back to the original table again, now 68 across A 1, 68 minus what is my G bar, G bar is 70, 68 what do you require to do 68 minus A l bar, this is the value. So, what is your A l bar I am just taking few first one from the first this column only I am only taking, so this one will be A 1 bar this minus A 2 bar then 73 minus A 3 bar, so this sense you have to make.

Now, again go to this what is what where we have calculated G bar A l bar A 1, A l bar is 71.80, so 68 minus 71.80 what will be this value is it not minus 0.80 then 85 minus 82.60 is it not 2.40 then 73 minus our 73.60 this minus 0.60. So, what I mean to say here then I mean to say that for each of the observations you have already want in table, now the mean of A 1 bar that is A 1 bar, A 2 bar, A 3 bar you have computed you subtract each of the cell across A 1 by A 1 bar. Similarly, A 2 cells by A 2 bar and A 3 cells by A

3 bar, whatever you will get will be the error of quantity correct, so once you know the error you have to go for certain.

(Refer Slide Time: 19:49)



95% CI for pair-wise comparisons

$$\bar{x}_{l.} - \bar{x}_{m.} - q_{\alpha,a,N-L}\sqrt{MSE/n} \le \mu_l - \mu_m \le \bar{x}_{l.} - \bar{x}_{m.} + q_{\alpha,a,N-L}\sqrt{MSE/n}$$

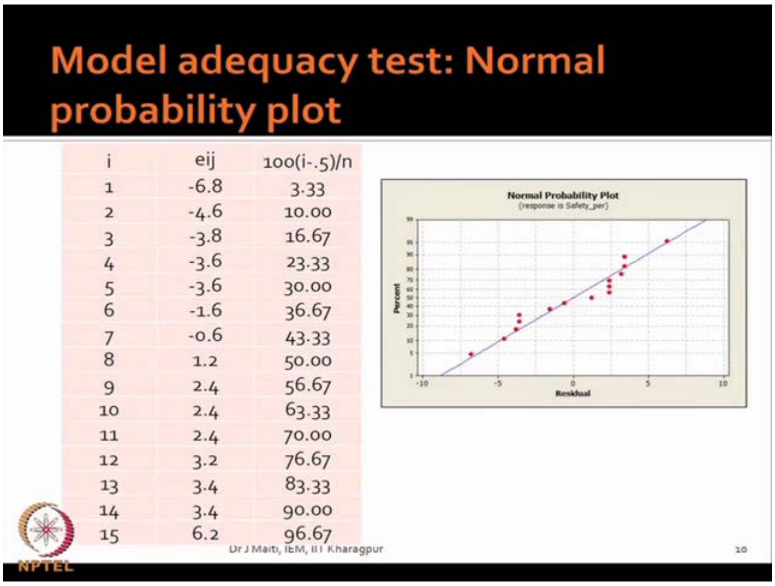
Difference	Lower	Center	Upper
A2 – A1	3.896	10.800	17.704
A3 – A1	-5.104	1.800	8.704
A3 – A2	-15.904	-9.000	-2.096

Dr J Maiti, IEM, IIT Kharagpur

9

Adequacy test as well as your...

(Refer Slide Time: 19:54)



Model adequacy test: Normal probability plot

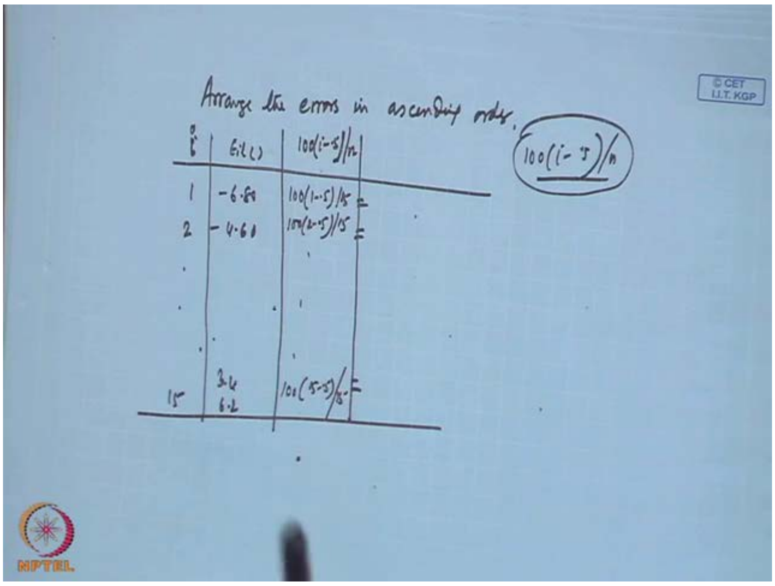
i	eij	100(i-.5)/n
1	-6.8	3.33
2	-4.6	10.00
3	-3.8	16.67
4	-3.6	23.33
5	-3.6	30.00
6	-1.6	36.67
7	-0.6	43.33
8	1.2	50.00
9	2.4	56.67
10	2.4	63.33
11	2.4	70.00
12	3.2	76.67
13	3.4	83.33
14	3.4	90.00
15	6.2	96.67

Dr J Maiti, IEM, IIT Kharagpur

10

Normality test and other test are there suppose I want to do the normality test, this one is known as model adequacy test what we will do normality test, how do you go about it because you have to do it from the errors.

(Refer Slide Time: 20:12)



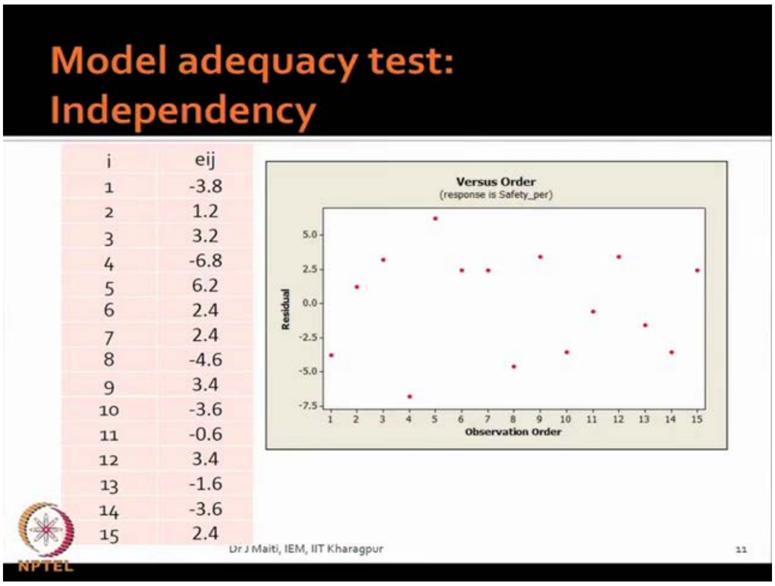
Arrange the errors in ascending order.

i	E_{iL} ( )	$100(i-.5)/n$
1	-6.80	$100(1-.5)/15 =$
2	-4.60	$100(2-.5)/15 =$
.	.	.
.	.	.
.	.	.
15	3.4 6.2	$100(15-.5)/15 =$

$\boxed{100(i-.5)/n}$

So, first step is arrange the errors in ascending order, so suppose in this case what you have you have 1, 2 like this 15 data points then your errors are coming something like this E i L. But, you are arranging them in ascending order, so you will be getting some values when you arrange, here you will be getting minus 6.80 minus 4.60 like this 3.4, 6.2 like this you will be getting. Then what you require to do you have to find out the probability value using this that is integral probability value again i what is to be written here 100 i minus 0.5 by your n, 100 into i minus 0.5 by n. So, it is 100 into 1 minus 0.5 by 15, 100, 2 minus 0.5 by 15 in this manner you will be writing last one 100, 15 minus 0.5 by 15, what value you will get this value you will be getting.

(Refer Slide Time: 21:45)



Model adequacy test: Independency

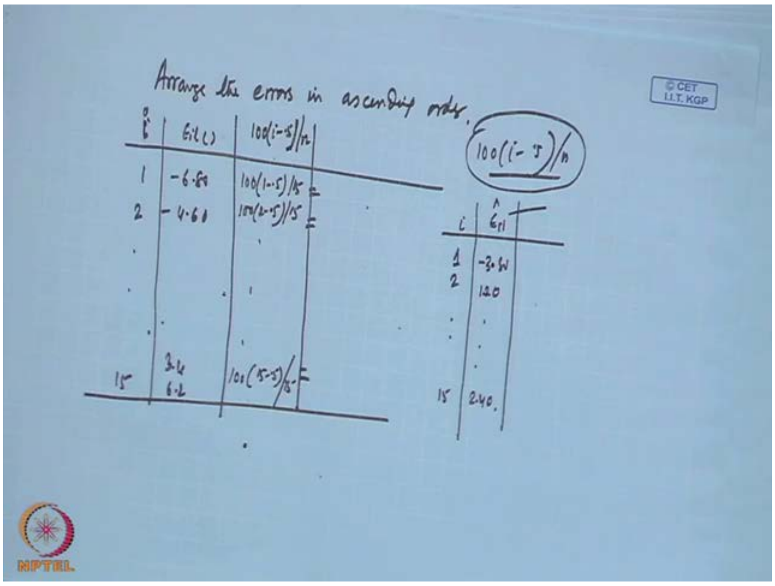
i	eij
1	-3.8
2	1.2
3	3.2
4	-6.8
5	6.2
6	2.4
7	2.4
8	-4.6
9	3.4
10	-3.6
11	-0.6
12	3.4
13	-1.6
14	-3.6
15	2.4

Dr J Maiti, IEM, IIT Kharagpur

11

So, first one 100 i minus 0.5 by n 3.33, so you must have e values that error values you arrange then in ascending order you have number of observations in serial it is arranged. Now, you are calculating this, so this one of the way of finding the probability values this in terms of percentage and then you plot what is a plot, plot is percent probability and then residual value this value. So, first residual is your minus 6, 6.8 probability value is 3.33 something it is coming somewhere here, so once you get like this then you draw a straight line. If your, if you get s good straight line in a sense the different error points will not differ much from the straight line, we can say that normality assumption is valid.

(Refer Slide Time: 22:49)



Arrange the errors in ascending order.

i	$e_{i()}$	$100(i-.5)/n$
1	-6.80	$100(1-.5)/15 =$
2	-4.60	$100(2-.5)/15 =$
.		.
.		.
.		.
15	3.4 6.2	$100(15-.5)/15 =$

$\boxed{100((i-.5)/n}$

i	$\hat{e}_{i1}$
1	-3.80
2	1.20
.	.
.	.
15	2.40.

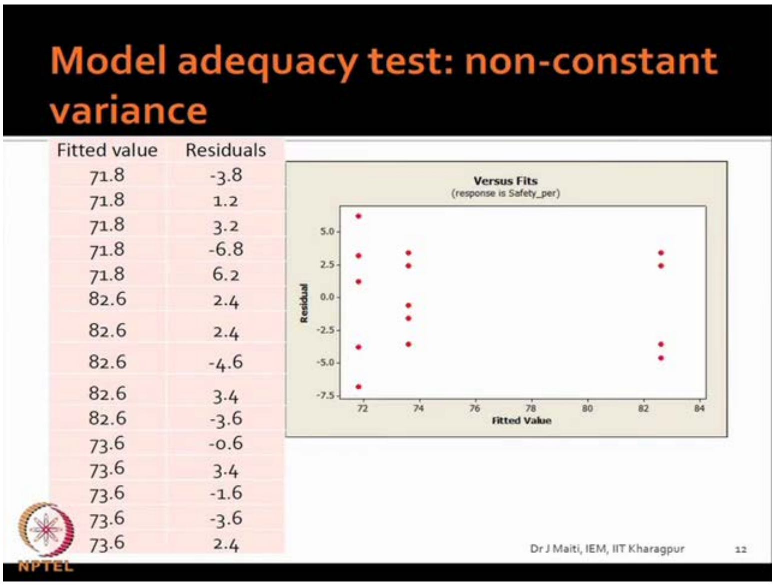
© CET I.I.T. KGP

Now, you may not be interested to use this 100 into i minus 0.5 by n you may go for some other formulation is there different people have different, but this is widely used one we are using this. So, this is the first what I can say the adequacy test, so please keep in mind what way we are proceeding we have data you have to first test hypothesis using the one way in your procedure S S T, S S A and S S E. Once the test is over and you are saying that there is difference then go for different parameters calculate all that parameters tau 1, tau 2, tau 3 or tau up to L.

Then compute the error also and using turkey test or that confidence interval use this also do the hypothesis testing for the parameters. So, they are significant or not or whether the difference between the parameters values are significant or not that is not of sufficient you have to do this adequacy test also, under adequacy test first point is normal probability plot. Now, come to the second point independency what we said we said that our, this observation they will be independent there will be no correlation between one observation to another observation.

That will be tested through if you plot the residuals of the errors again you are putting observation order in this case in this case our observation order is 1, 2 like this 15 and you are writing. Here, epsilon i l estimate the way it is estimated not the arranged one, it is not arranged in ascending or descending order it is the way your first observation is this what is the value second observation is, this what is the value. So, then first observation value will be 3.80 second one 1.20, so like this it will go to the 15 one 2.40 this one you arrange d for normal probability plotting, but here you do not require to arrange here.

(Refer Slide Time: 25:14)



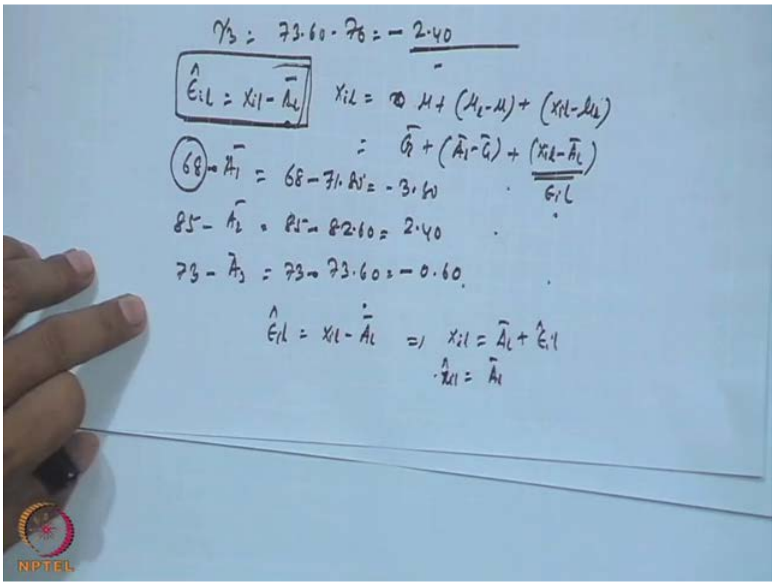
Model adequacy test: non-constant variance

Fitted value	Residuals
71.8	-3.8
71.8	1.2
71.8	3.2
71.8	-6.8
71.8	6.2
82.6	2.4
82.6	2.4
82.6	-4.6
82.6	3.4
82.6	-3.6
73.6	-0.6
73.6	3.4
73.6	-1.6
73.6	-3.6
73.6	2.4

Dr J Maiti, IEM, IIT Kharagpur 12

You see the slide here, so these are the observation numbers and this is basically the residual value and if you plot and you must not get any clear. So, we can say there is no serial correlation between one observation to another third one what you have to test that non constant variants we said that the variants is equal across the population. So, here what you require to do you require to find out the fitted values and residuals and what are the fitted values here what will be the fitted value.

(Refer Slide Time: 26:07)



$\tau_3 = 73.60 - 76 = -2.40$

$\boxed{\hat{\epsilon}_{il} = x_{il} - \bar{A}_l}$ $\quad x_{il} = \mu + (\mu_l - \mu) + (x_{il} - \mu_l)$

$= \bar{G} + (\bar{A}_l - \bar{G}) + \underbrace{(x_{il} - \bar{A}_l)}_{\epsilon_{il}}$

$68 - \bar{A}_1 = 68 - 71.80 = -3.80$

$85 - \bar{A}_2 = 85 - 82.60 = 2.40$

$73 - \bar{A}_3 = 73 - 73.60 = -0.60$

$\hat{\epsilon}_{il} = x_{il} - \bar{A}_l \Rightarrow x_{il} = \bar{A}_l + \hat{\epsilon}_{il}$

$\hat{x}_{il} = \bar{A}_l$

So, you see what we have written here that epsilon i L is x i l minus A L bar, so that means you have written like this that x i l equal to A L bar plus epsilon i L fitted value means without error. So, that means I can say x i l fitted value is A L bar, so as a result what will happen you will find that for the first for population one that is A 1 for the first five values the it is replaced by the average this average across A 1. Similarly, average across A 2 second five values, third five values is average across A 3 this is what are the fitted values and then residuals you have already computed across everything.

Now, you fit or plot against residuals to fitted value, here are you getting any difference in variability I think this one first one little difference is there because the mean value 71.80 and this across this mean the variability is little higher compared to the other two. But, theses two are almost equal, but we have tested these you have to test through some statistics that test I told you in earlier.

So, here it is showing that that the population one that means the level one A 1 variability is little more it is apparently like this it may not be true because this is A this is basically a plot where the scale differences will make changes. So, I think there may not be different, but it should be tested, but we have not tested in this example, but earlier examples we have tested, so if you if there is departure from equal variability then it will be reflected in the fitted value versus the residual plot, so this is our one way ANOVA

Student: Sir, if there is a difference in the assumptions as it is coming from equal variability so can we use an ANOVA.

We can use, so you are saying that there is violation of equality of variances yes you can still use ANOVA that ANOVA statistics for ANOVA procedure is robust procedure. But, again what I can say is that in that case it is advisable to take equal sample size across the populations, if you do not take equal sample size there will be little problem but you have to be very much.

Student: But, these assumptions are very much.

Vital assumptions.

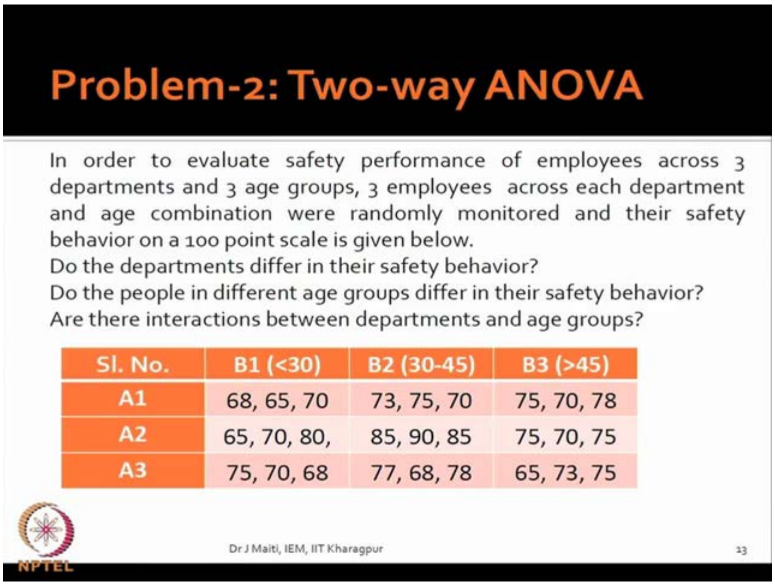
Student: Vital yes.

Vital as well as complicated many most of the times in a real life situation it is difficult to get normality you may get, but the equal variants part it is difficult. So, under such conditions what I suggest that if possible you go for equal sample size and increase the sample size and second is that other way you transform the variables there are different types of transformation. For example, power transformation is there if there is normality problem then box cox transformation is there, so most of the time if it is not homogeneous, so normality assumption is violated.

So, if it is homogeneous and if it normal traditional it can be it is homogeneous also the homo part another thing is that you before doing ANOVA why cannot you remove the outliers because the variability will be very much affected by the outliers. So, you remove the outliers if you remove the outliers it may, so happen that your variability component will be within the control this way you please try.

But, again I am telling you I have done three four different analysis concerning the equality of variances particularly in the multi variate domain very difficult we are not getting equal co variant matrix. Then I searched what to do there are many statistics which are basically robust against this type of problem you have to search appropriate statistics and use.

(Refer Slide Time: 31:20)



Problem-2: Two-way ANOVA

In order to evaluate safety performance of employees across 3 departments and 3 age groups, 3 employees across each department and age combination were randomly monitored and their safety behavior on a 100 point scale is given below.
Do the departments differ in their safety behavior?
Do the people in different age groups differ in their safety behavior?
Are there interactions between departments and age groups?

Sl. No.	B1 (<30)	B2 (30-45)	B3 (>45)
A1	68, 65, 70	73, 75, 70	75, 70, 78
A2	65, 70, 80,	85, 90, 85	75, 70, 75
A3	75, 70, 68	77, 68, 78	65, 73, 75

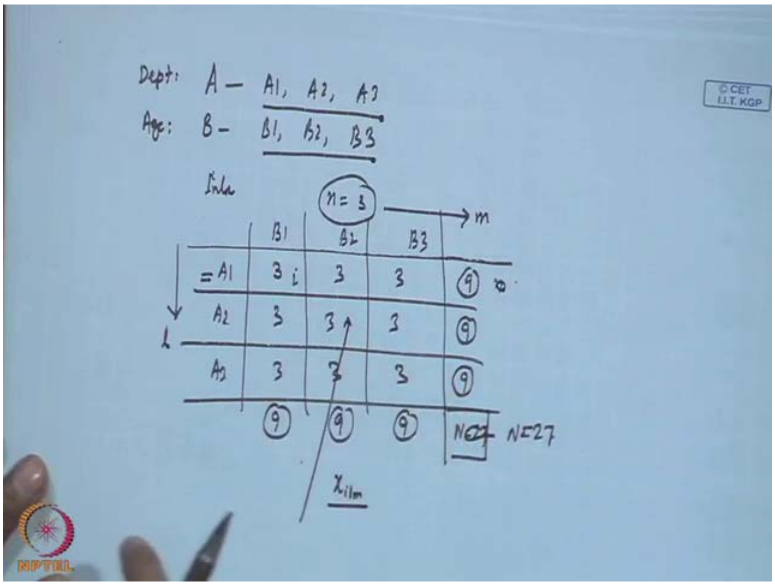
Dr J Maiti, IEM, IIT Kharagpur

13

Problem 2, two way ANOVA I think last class we discussed MANOVA, but before MANOVA I have given you some idea then when two factors are governing the behavior of the system then two independent factors or dependent factors. So, depending on the nature of the relationship between the factors you may find out that the behavior of the system will be different now here in the same example, but the data are hypothetical data. So, you may not get the exact one to one relationship between the department affects, here in this problem with the problem given in one that may not be correct, but let us see that it is a separate problem.

Now, that in order to safety performance as I told you three departments and three age groups are considered that means what is our assumption. Here, is that the three departments are three population type population or levels and different age groups also have different levels, different levels that these are the different levels. So, different populations that sense, so hence forth as two factors are coming we will not use the word population frequently, we will use the word factors, so department is one factor that is A and age is another factor that is B.

(Refer Slide Time: 32:59)



Dept: A – A1, A2, A3

Age: B – B1, B2, B3

$n = 3$ → m

	B1	B2	B3	
A1	3	3	3	9
A2	3	3	3	9
A3	3	3	3	9
	9	9	9	N=27

l

x_{ilm}

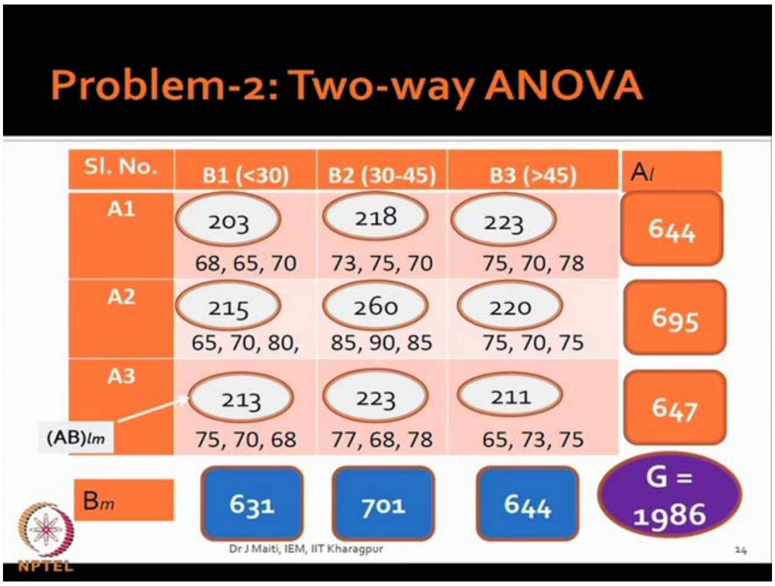
Now, department has A factor, has three levels A 1, A 2 and A 3 and B factor which is the age factor which is also having three levels in this example. So, that means we are saying A means department B is the age what do you want to say we are trying to establish that there is it that there is difference across departments. In the safety behavior, is there difference between age groups is there interaction between age and department interaction between age and department that is clearly written.

Here, do the departments differ in their safety behavior that is our first question, do the people in different age groups differ in their safety behavior are there interaction between departments and age groups in terms of safety behavior. Now, if your query is only the first one and you do not consider that age is an important factor then it only one way ANOVA, what we have done earlier as we are saying that it is not only the department is talking about the working condition. But, the people behavior, people expertise, people that thinking that is not coming under department we want to capture these through different age groups, so as safety is an issue of the unsafe acts and the unsafe conditions point of view.

So, first the department one we are treating it as unsafe condition point of view and age factor we are considering that individual influence point of view. Now, suppose three data points equal sample sizes, so it is again suggested that it is better you go for equal sample sizes whenever possible. So, here n equal to 3 and our three departments A 1, A 2 and 3 and three age groups B 1, B 2 and B 3, so here three data points 3, 3 means across A 1 data point 9, 3, 3, 3 this also 9, 3, 3, 3 this also 9. Similarly, across age group one age group, two age group, 3, 9 data points, so that means in this is in if I say that with respect to this is 9 like this some notation we can use.

Here, then the gross is the total capital is your 27 this is the latest as per as the by the number of observations point of view is the case and you can see that what are the different values. Suppose x if you want to use the notation like this, for departments will be using l for age groups will be using m and for observation will be using i. So, any observation, here it can be denoted by A x i l m, any observation can be denoted by A x i L m, so is this is A 1, B 1, so x 1 1 1 is 68, x 1 1 2 is 65, x 1 1 3 73 like this.

(Refer Slide Time: 36:47)



Problem-2: Two-way ANOVA

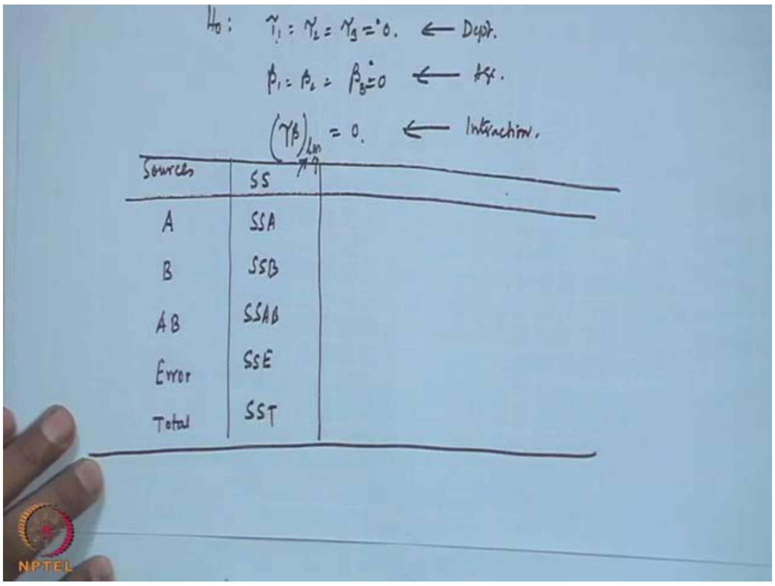
Sl. No.	B1 (<30)	B2 (30-45)	B3 (>45)	A_l
A1	203 68, 65, 70	218 73, 75, 70	223 75, 70, 78	644
A2	215 65, 70, 80,	260 85, 90, 85	220 75, 70, 75	695
A3	213 75, 70, 68	223 77, 68, 78	211 65, 73, 75	647
B_m	631	701	644	G = 1986

Dr J Maiti, IEM, IIT Kharagpur

14

What we want we want to first see that is there any effect of the departments, effect of the age groups and interaction effects, so we will go by hypothesis testing what is your hypothesis, here in this example.

(Refer Slide Time: 37:10)



H_0: $\tau_1 = \tau_2 = \tau_3 = 0.$ ← Dept.

$\beta_1 = \beta_2 = \beta_3 = 0$ ← Age.

$(\tau\beta)_{lm} = 0.$ ← Interaction.

Sources	SS
A	SSA
B	SSB
AB	SSAB
Error	SSE
Total	SST

My hypothesis is that from the departments point of view if I say tau 1 equal to tau 2 equal to tau 3 equal to 0, I am not writing mu 1, mu 1 mu all those things then this is from department point of view. Similarly, from suppose beta 1 equal to beta 2 equal to beta 3 equal to 0 that is from age point of view, now similarly what will happen basically tau beta, tau beta. Now, I am writing, now l m, l stands for the department and m stands for the age that will also become 0, now the combination will be there tau l and all those things.

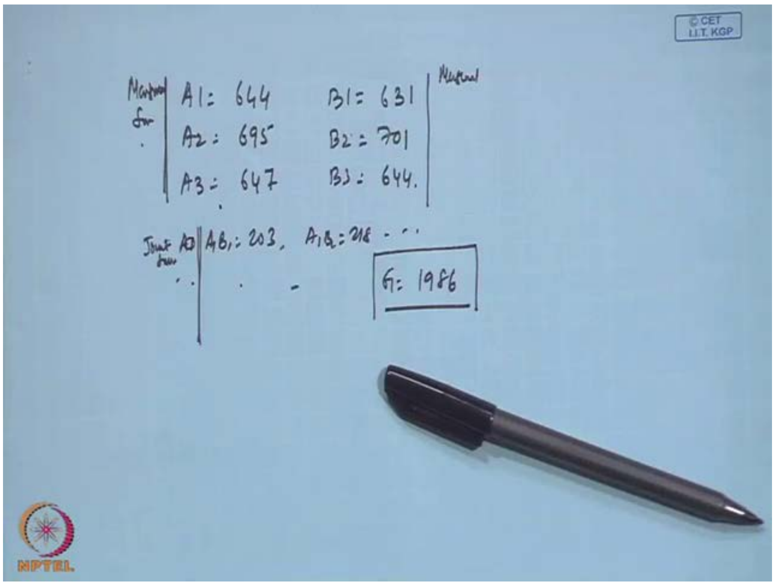
So, why is this coming, so this coming because of interaction point of view if this is the case then what are the sources of variation, first one is department which is A second one is age which is B third one interaction which is A B, fourth one is error, fifth one is that this is the total one this four sum is the total. So, this is our sources of variations, so what is required to compute you require to compute S S, S S means sum square, sum square A, sum square B sum square A B sum square error, sum square total.

So, how do compute this, what will be the computation of your S S A you just see this table it is not changing this way.

Student: No sir, it is coming.

I am seeing that when you no problem, so see that what is the total what is the total across A 1 total mean 68 plus 65 plus 70 all values this is 644, for A 2 695, for A 3 it is 647, so we are denoting it as A 1, A 2 and A 3.

(Refer Slide Time: 39:48)



Marginal for

A_1: 644 B_1 = 631

A_2: 695 B_2 = 701

A_3: 647 B_3: 644.

Marginal

Joint AB for

A_1B_1: 203, A_1B_2: 218 . . .

G: 1986

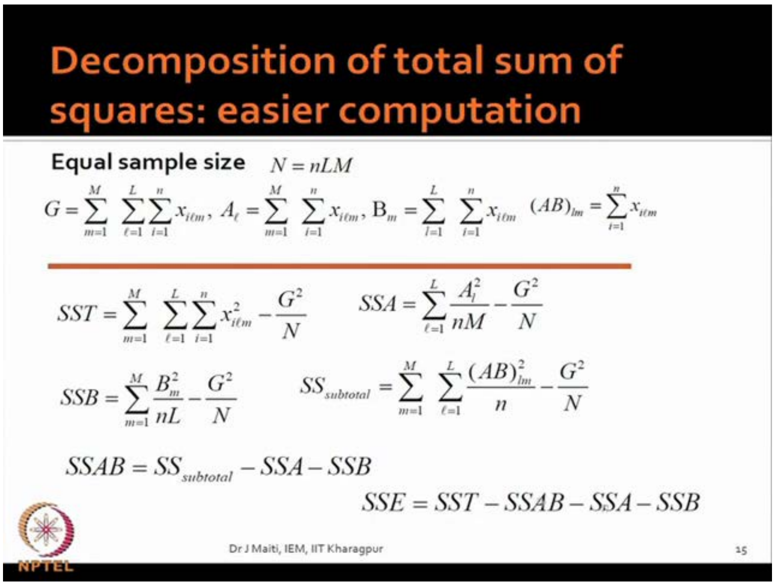
So, write down A 1 equal to 644, A 2 equal to 695, A 3 equal to 647, similarly you see across the column when you make the total you will be getting the that column total which are basically for different age group. So, you will be getting B 1 equal to 631, B 2 equal to 701, B 3 equal to 644 another thing which is also equally important, here we want to see the interactions, so the interactions each of sale is the interaction part.

Now, you are making the sum total of each of the sales that is A B part, so this one if I say sum A 1, B 1 is 203, so if I write like this A 1, A 1 B 1, 203 similarly A 1 B 2, 218 these are all sum. So, you can you can see that 203, 218, 223, 215, 260 all those things are coming under the total across the total in each cell, so that we are denoting by A B l m A, so A B 1 1, now A B 1 2, A B 1 3 like this. So, how many sums your getting one is the if I say this is the marginal sum across department, this is marginal sum across department, this one is marginal sum across age group and then these are basically joint sums.

Joint sum means, basically these are the combination each combination what is the sum value and then you also required to compute the grand total in this case our grand total is 1986. So, when you required to go for two way ANOVA you will be getting a table which is known as contingency table, it is a little it is contingency table looking.

So, in the contingency table there are marginal counts across rows across columns and joint counts, but here our value not counts value, but our values are basically those continuous values but you are getting values. So, we will not be using contingency table word we will use that this is our data set, now you are calculating three four things initially one is across rows the total across column, the total across sales, the total and then the grand total these things will be used to compute S S T, S S A, S S B all those things.

(Refer Slide Time: 43:15)



Decomposition of total sum of squares: easier computation

Equal sample size $N = nLM$

$$G = \sum_{m=1}^{M} \sum_{\ell=1}^{L} \sum_{i=1}^{n} x_{i\ell m},\ A_\ell = \sum_{m=1}^{M} \sum_{i=1}^{n} x_{i\ell m},\ \mathrm{B}_m = \sum_{l=1}^{L} \sum_{i=1}^{n} x_{i\ell m}\ \ (AB)_{lm} = \sum_{i=1}^{n} x_{i\ell m}$$

$$SST = \sum_{m=1}^{M} \sum_{\ell=1}^{L} \sum_{i=1}^{n} x_{i\ell m}^2 - \frac{G^2}{N} \qquad SSA = \sum_{\ell=1}^{L} \frac{A_l^2}{nM} - \frac{G^2}{N}$$

$$SSB = \sum_{m=1}^{M} \frac{B_m^2}{nL} - \frac{G^2}{N} \qquad SS_{subtotal} = \sum_{m=1}^{M} \sum_{\ell=1}^{L} \frac{(AB)_{lm}^2}{n} - \frac{G^2}{N}$$

$$SSAB = SS_{subtotal} - SSA - SSB$$

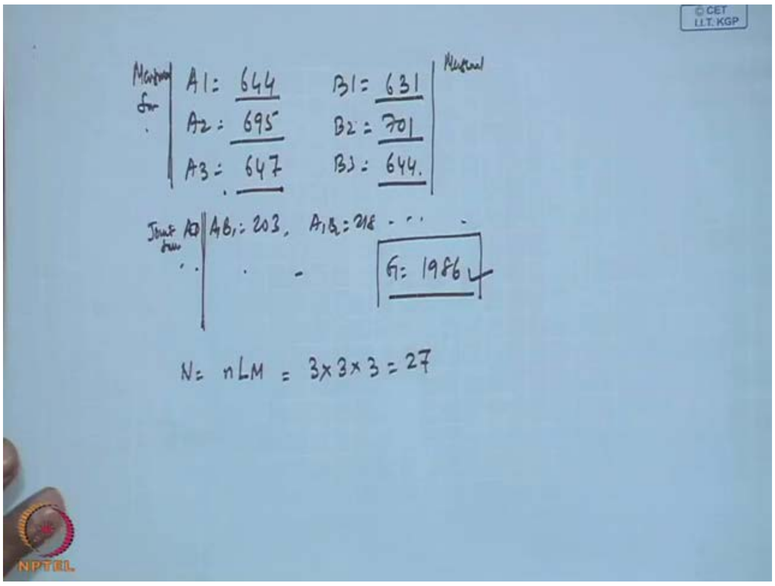
$$SSE = SST - SSAB - SSA - SSB$$

Dr J Maiti, IEM, IIT Kharagpur

15

This is our computational formulas from equal sample size point of view you see first what is the total observation, total observation here.

(Refer Slide Time: 43:32)



Marginal for
Marginal
Joint AB

A1: 644 B1= 631
A2: 695 B2 = 701
A3: 647 B3: 644.

A_1B_1: 203, A_1B_2 = 218 . . .

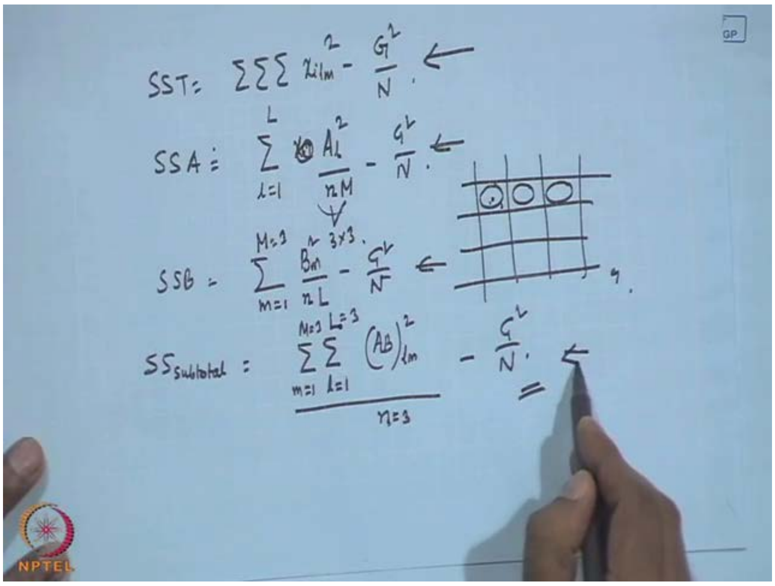
G: 1986

N= nLM = 3x3x3 = 27

In this example N equal to n L into M where our N is 3, L is 3 and M is 3, so this is 27 and that you have seen earlier and then second quantity what we are computing we are computing the grand total G. So, that is going through all the sales, so as a result it is in all the values in each cell so I equal to 1 to n this is talking about all the values in one cell way and L equal to 1 2 for the departments or a factors M equal to factors this is the grand total.

Now, grand total values we got it is 18, 19, 86, now A L 3, A L values you have already seen that is 644, 695, 647 then you got three values across age that is B M that is 631, 701, 644 and you have a large number of values where there is 9 values where is there are 3 cross 3 table. So, 9 values for joint A B jointly what is happening what are the values that you will be getting the addition, some of these values you are getting.

(Refer Slide Time: 45:04)



$$SST = \sum\sum\sum x_{ilm}^2 - \frac{G^2}{N}.$$

$$SSA = \sum_{l=1}^{L} \frac{A_l^2}{nM} - \frac{G^2}{N}.$$

$$SSB = \sum_{m=1}^{M=3} \frac{B_m^2}{nL} - \frac{G^2}{N}$$

3×3.

9.

$$SS_{subtotal} = \frac{\sum_{m=1}^{M=3}\sum_{l=1}^{L=3} (AB)_{lm}^2}{n=3} - \frac{G^2}{N}.$$

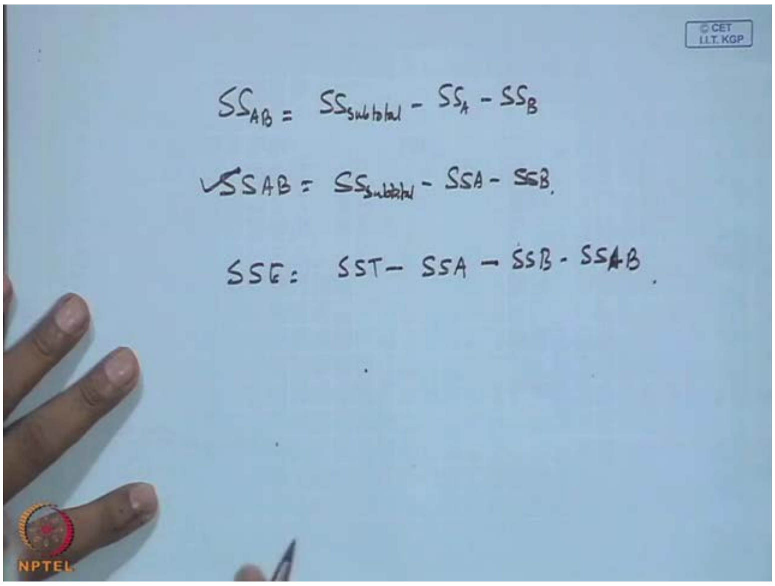
So then what will be your S S T calculation, S S T calculation will be S S T is 3 sum x i l m this square minus G square by N, so when you talk about S S A, S S A will be you are considering only the rows, so l equal to 1 to l. That will be your x i, no we have already created A L square by n M minus G square by n because again against each row you will be finding out that we are not considering the B factor affect. So, ultimately there are three cross three nine observations, so here 9 observations 3 cross 3, 9 observations, so what will be our S S B, S S B across m equal to 1 to M then that is B m square by n will be there.

But, m is varying, so l will be written here and minus G square by N, now we will create another thing which is known as S S sub total, S S sub total if you see that I told you that 3 cross 3. Here, the sum of all values in this cell sum of like this 9 values are there, so these 9 values will be coming from l equal to 1 to L and m equal to 1 to M where L equal to N equal to 3, here and every where this is equal to 3. So, that 9 values sum values these values we have given a name A B l m you have seen earlier this value I have given this name, so this square and it is basically addition of three values.

So, divided by n that equal to 3 minus G square by N you want to see this and showing this one you see what is will explain with respect to this. So, what we are saying that S S T first one is S S T 68, 65, 70 all values you square them and add minus this G square by N this is the formula all values in the table these are squared and then add it then minus this. Now, for the second S S A see what we have taken we have taken the sum across the rows 644, 695, 647 and these values you are squaring first then n dividing it by that here 3 plus 3 plus 3, 9 values are there, so it is the second one.

So, n M divided by N M, so similarly you will be doing this for across column just see across column 631, 701, 644, now subtotal means subtotal means this is subtotal 203 A B l m 203, 218, 223, 215, 260 like this. So, you are squaring these, these, these all then making total then dividing it by this the number of observations and then following like this. So, you will be getting subtotal, but once you get subtotal you will you are in a position to calculate.

(Refer Slide Time: 49:12)



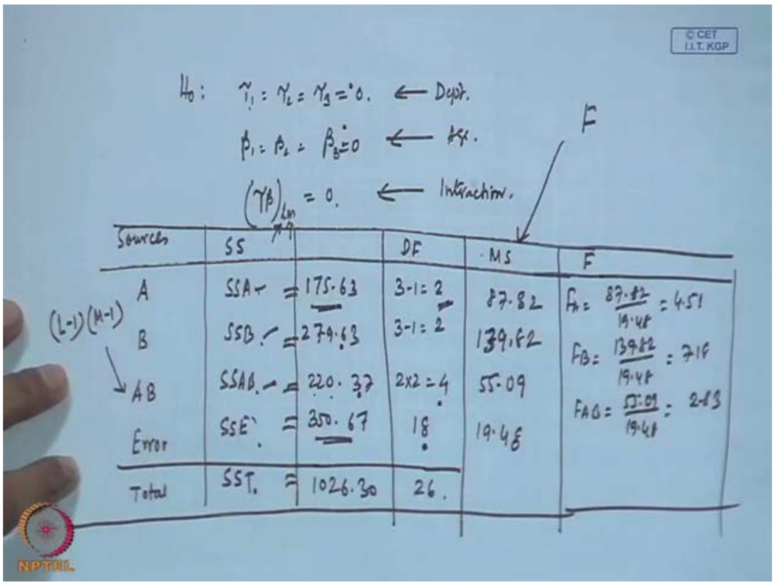
$SS_{AB} = SS_{subtotal} - SS_A - SS_B$

✓$SSAB = SS_{subtotal} - SSA - SSB.$

$SSE: SST - SSA - SSB - SSAB$.

The interactions parameters S S A B, S S A B will be S S subtotal minus S S A minus S S B, so in this way you can write or the way we have written first S S A B equal to S S subtotal minus S S A minus S S B. So, what are the items you have now you have S S T, you have S S A, you have S S B, you have S S subtotal and you have S S A B what is required to be calculated, now S S E, S S E will be S S T minus S S A minus S S B minus S S A B, because you see.

(Refer Slide Time: 50:17)



© CET I.I.T. KGP

$H_0: \tau_1 = \tau_2 = \tau_3 = 0. \leftarrow$ Dept.

$\beta_1 = \beta_2 = \beta_3 = 0 \leftarrow$ Exp.

$(\tau\beta)_{lm} = 0. \leftarrow$ Interaction.

F

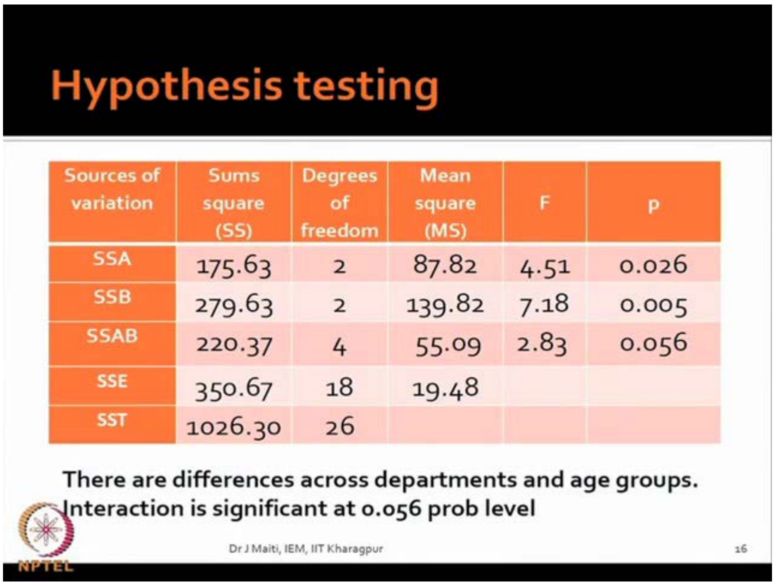
Sources	SS		DF	MS	F
A	SSA	= 175.63	3-1 = 2	87.82	$F_A = \frac{87.82}{19.48} = 4.51$
B	SSB	= 279.63	3-1 = 2	139.82	$F_B = \frac{139.82}{19.48} = 7.18$
AB	SSAB.	= 220.37	2x2 = 4	55.09	$F_{AB} = \frac{55.09}{19.48} = 2.83$
Error	SSE.	= 350.67	18	19.48	
Total	SST.	= 1026.30	26.		

$(l-1)(m-1)$

The total is S S A, S S B, S S A B, S S E, S S T, so S S E will be S S T minus this, minus this, minus this, so this is the calculation part decomposition of total sum square into individual sum squares those individuals are the sources. We have computed this based on the computation you got the values like this is S S A equals to 175.63, S S B 279.63, S S C 220.37, S S E 350.67 and S S T 1026.30 these are the S S values. Now, what will be your degrees of freedom, here what is the total observation 27, so 27 minus 1 that is 26 how many levels for A 3, so 3 minus 1, 2 how many levels for B 3, 3 minus 1, 2 what will be the A B case 3 minus 1 into 3 minus 1 that means A Bs case.

Here, A has l minus 1 degree of freedom B has m minus 1 degrees of freedom multiplication, so 2 cross 2 equal to 4 the rest 26 minus 4 minus 2 minus 2 that mean minus 26 minus 8 that is 18 is the error degrees of freedom. Now, what will be your M S then M S will be 175 by 2 that is 87.82, 279.63 by 2 that is 139.82 then 220.37 by 4 this will be 55.09, 350.67 divided by 18 this is your 19.48. Then what you require to know, you require to know F values your F A will be M S A by M S E, so 87.82 by 19.48 this value will be 18 in 4.51, then f B 139.82 by 19.48 this value is 7.18 then F A B which is 55.09 by 19.48 which is our 2.83. So, your ANOVA table is ready what you require, now you require to find out tabulated value let some value alpha value you take 0.05 all those things.

(Refer Slide Time: 53:39)



Hypothesis testing

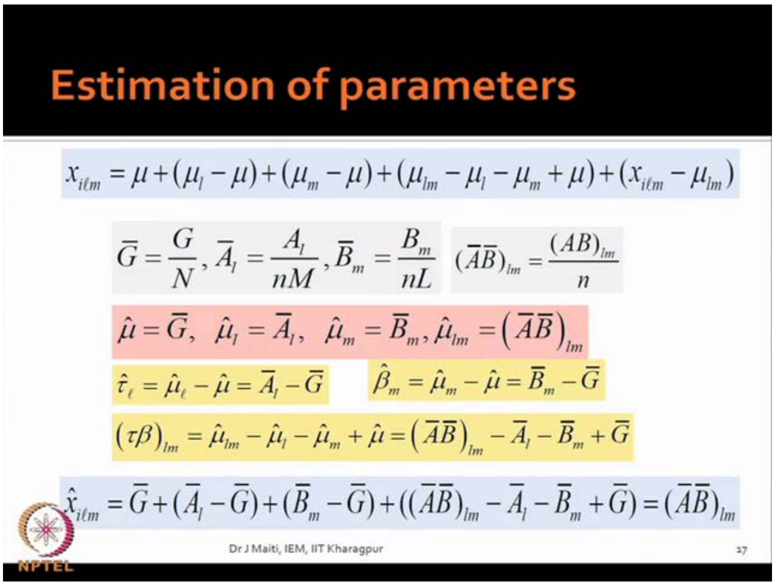
Sources of variation	Sums square (SS)	Degrees of freedom	Mean square (MS)	F	p
SSA	175.63	2	87.82	4.51	0.026
SSB	279.63	2	139.82	7.18	0.005
SSAB	220.37	4	55.09	2.83	0.056
SSE	350.67	18	19.48		
SST	1026.30	26			

There are differences across departments and age groups. Interaction is significant at 0.056 prob level

Dr J Maiti, IEM, IIT Kharagpur 16

Now, you see this we have identified what is the P value for a figure 4.57 with degrees of freedom 2 and 18 this is point 0 to 6 it is less than 0.05 second one is also less than 0.05, but third one is not less than 0.05. So, if you consider alpha equal to 0.05 that means 95 percent confidence or 0.05 percent significance level then your A and B effect are significant where interaction is effects are not significant. But, these values very close to 0.05, 0.05, 6 is very close to 0.05. So, what I mean to say then you may also consider it significant the result is there a difference across departments in different age groups interaction is significant up to probability levels. So, we can say also interaction effect is there now once this test is over now what is your next you have to test the parameters.

(Refer Slide Time: 54:52)



Estimation of parameters

$$x_{i\ell m} = \mu + (\mu_l - \mu) + (\mu_m - \mu) + (\mu_{lm} - \mu_l - \mu_m + \mu) + (x_{i\ell m} - \mu_{lm})$$

$$\bar{G} = \frac{G}{N}, \bar{A}_l = \frac{A_l}{nM}, \bar{B}_m = \frac{B_m}{nL} \quad (\bar{A}\bar{B})_{lm} = \frac{(AB)_{lm}}{n}$$

$$\hat{\mu} = \bar{G}, \quad \hat{\mu}_l = \bar{A}_l, \quad \hat{\mu}_m = \bar{B}_m, \hat{\mu}_{lm} = \left(\bar{A}\bar{B}\right)_{lm}$$

$$\hat{\tau}_\ell = \hat{\mu}_\ell - \hat{\mu} = \bar{A}_l - \bar{G} \qquad \hat{\beta}_m = \hat{\mu}_m - \hat{\mu} = \bar{B}_m - \bar{G}$$

$$(\tau\beta)_{lm} = \hat{\mu}_{lm} - \hat{\mu}_l - \hat{\mu}_m + \hat{\mu} = \left(\bar{A}\bar{B}\right)_{lm} - \bar{A}_l - \bar{B}_m + \bar{G}$$

$$\hat{x}_{i\ell m} = \bar{G} + (\bar{A}_l - \bar{G}) + (\bar{B}_m - \bar{G}) + ((\bar{A}\bar{B})_{lm} - \bar{A}_l - \bar{B}_m + \bar{G}) = (\bar{A}\bar{B})_{lm}$$

Dr J Maiti, IEM, IIT Kharagpur

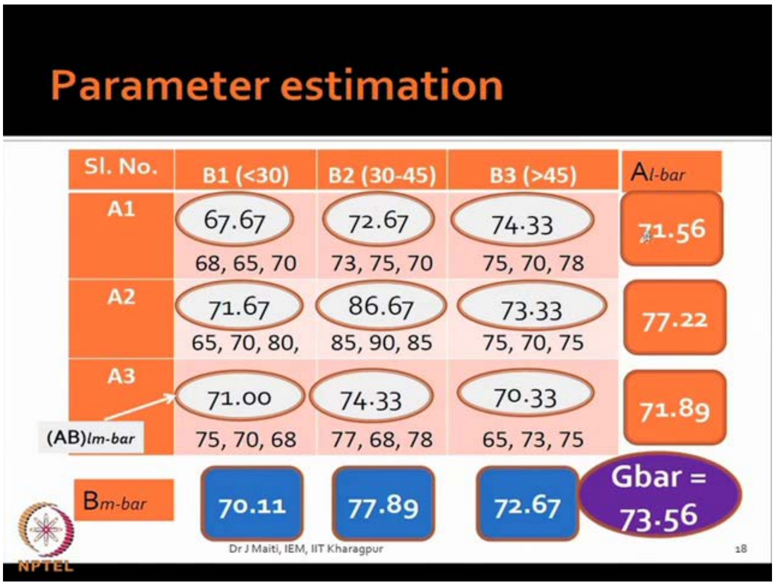
17

These are the formulas for estimation of parameter you see, here again we are partitioning the observation i l m observation mu class mu l minus mu this is the A effect mu m minus mu that is the B effect. Then mu l m minus mu l minus mu m plus mu this one is we have taken, here mu l, so you are subtracting mu l, here mu A minus mu plus mu this mu this mu plus.

So, it is cancelling finally, x l A x i l m minus mu l m because one mu l m is added, here and this is the partition, so this first one then mean second one is A effect, say third one is B effect, fourth one is A B that is interaction effect last one is error. So, this one we are writing in the same manner, so grand mean, now G by n, now A effect that A L bar A L by n m, B M bar B M by N L then A B bar l m that is A B l m by n then mu is nothing but G bar mu l is nothing but A L bar mu m. But, B M bar mu l m is nothing but A B l m bar then tau l will be that mu l minus mu l cap minus mu cap is the tau l cap, so that is A L bar minus G bar.

Similarly, beta m is also you are calculating, similarly the interaction between you are calculating and then what will be the predicted value of each observations this will be except the e i l component. So, this component is not there now if when we do you manipulate this one what you will find, you will find out that this one is nothing but A B bar l m this is the joint that sale average are the predicted values getting it.

(Refer Slide Time: 57:22)



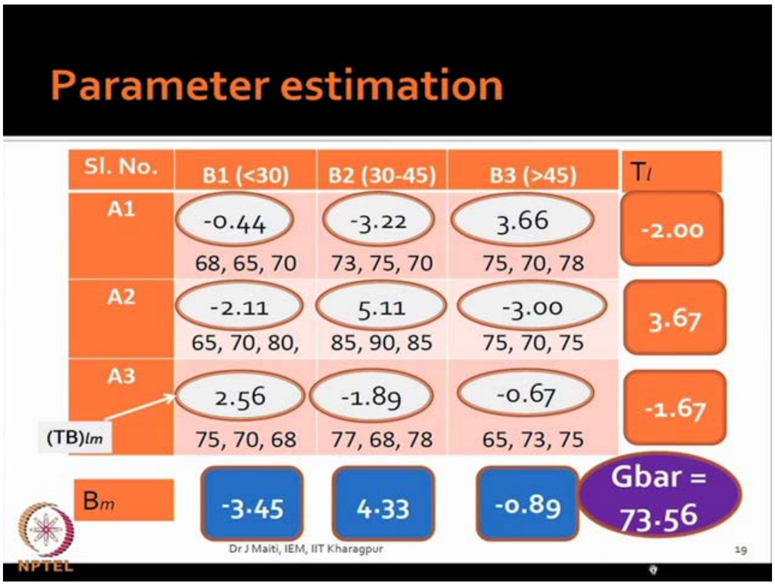
Parameter estimation

Sl. No.	B1 (<30)	B2 (30-45)	B3 (>45)	$A_{l\text{-}bar}$
A1	67.67 68, 65, 70	72.67 73, 75, 70	74.33 75, 70, 78	71.56
A2	71.67 65, 70, 80,	86.67 85, 90, 85	73.33 75, 70, 75	77.22
A3 $(AB)_{lm\text{-}bar}$	71.00 75, 70, 68	74.33 77, 68, 78	70.33 65, 73, 75	71.89
$B_{m\text{-}bar}$	70.11	77.89	72.67	Gbar = 73.56

Dr J Maiti, IEM, IIT Kharagpur 18

So, you see that this calculation we have done like this, so this is our, this is our 71.56 A L bar if you, if you subtract this minus you will be getting the tau 1, 77.22 minus this tau 2, 71.89 minus this tau 3. Similarly, beta 1, beta 2, beta 3; similarly you will be getting tau beta 1, 1 tau beta 1, 2 tau beta 1, 3 that means 67.67 minus this way you will be able to get.

(Refer Slide Time: 57:58)



Parameter estimation

Sl. No.	B1 (<30)	B2 (30-45)	B3 (>45)	T_l
A1	-0.44 68, 65, 70	-3.22 73, 75, 70	3.66 75, 70, 78	-2.00
A2	-2.11 65, 70, 80,	5.11 85, 90, 85	-3.00 75, 70, 75	3.67
A3 $(TB)_{lm}$	2.56 75, 70, 68	-1.89 77, 68, 78	-0.67 65, 73, 75	-1.67
B_m	-3.45	4.33	-0.89	Gbar = 73.56

Dr J Maiti, IEM, IIT Kharagpur 19

Now, see what is happening you are getting every where the tau l B m and tau B l these are the parameters apart from this what para you want. You want to calculate the error component also, error component is also you parameter I think will start again we will start again this is…

Thank you very much.