

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 – 15

Topic

Analysis of Variance (ANOVA)
(Contd.)

We will continue ANOVA for another 15, 20 minutes.

(Refer Slide Time: 00:23)

MANOVA

$H_0: \mu_1 = \mu_2 = \dots = \mu_L$
 $H_1: \mu_2 \neq \mu_m$

Null Hypothesis is rejected

$\hat{\mu}_1 = \bar{x}$
 $\hat{\mu}_2 = \bar{x}_2$
 $\hat{\mu}_2 = \hat{\mu}_2 - \hat{\mu}_1 = \bar{x}_2 - \bar{x}$

$\hat{\mu}_2 = \mu_2 - \mu_1$
point estimation

$\hat{\mu}_2 = \mu_2 - \mu_1$
Interval estimation

Point estimate

SECRET IIT KGP

Then we will discuss MANOVA. We have seen the hypothesis testing using ANOVA, where we have said that all the means are equal and alternative hypothesis at least one pair of means are not equal, and using F test, we either accept or reject the null hypothesis. For the problem, consider we find that the null hypothesis is rejected, so in our case the null hypothesis is rejected. Now, we will discuss two things, one is that when you talk about τ_L that is the population effect that is called for l-th population.

We want to compute it how to compute this τ_L , there will be two cases, one will be point estimation and second one will be interval estimation. What we have seen earlier for population mean and difference between two population mean cases, okay. So, in this case for ANOVA that grand mean estimate of grand mean is the grand sample means, then estimate of population mean is again the sample population mean and estimate of population effect is your estimate of that is the population mean minus grand mean, which is $\bar{x}_L - \bar{x}$.

So, this I can say that point estimation, this is your, I can say that point estimation, this is what your point estimation. Now, you want to find out the interval estimation what you will do, what you require to know like earlier.

(Refer Slide Time: 03:30)

$\tau_L = \bar{x}_L - \mu \sim \text{Random variable.}$

$\underline{MSE} = \frac{SSE}{N-L} = \hat{\sigma}^2$

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

$\bar{x}_L \sim N(\mu_L, \hat{\sigma}^2/n) = N(\mu_L, \frac{MSE}{n})$

$\hat{\sigma}^2 = MSE = N(\mu_L, \frac{MSE}{n}) \leftarrow$

$\Rightarrow \bar{x}_L - t_{(N-L)} \sqrt{\frac{MSE}{n}} < \mu_L < \bar{x}_L + t_{(N-L)} \sqrt{\frac{MSE}{n}}$

We can say that the τ_L which is your $\bar{x}_L - \bar{x}$, this is a random variable, the reason is both all are estimates only, so if you know a random variable and you know, you must know that distribution of that random variable. You must know the statistic and the sampling distribution of the statistic then only you will be able to find out the interval estimation. In this case, τ^{\wedge} or μ^{\wedge}_L we want to know the interval estimation. In ANOVA we have computed MSE, MSE is a SSE by its degrees of freedom that is N-L.

This is estimate of σ^2 , okay what is this σ^2 . You have recall the bullet test where we said that $\sigma_1^2 = \sigma_2^2$, all population variances are equal to σ^2 . So, this one is the estimate of σ^2 , so MSE. So we know MSE is known, now if I want to know what the distribution of \bar{x}_L , then we will say this will follow normal distribution with μ_L and that σ_L that is $\sigma \bar{x}_L^2$ this one and you can remember that this will be nothing but $\mu_L \sigma^2 n$. Now, σ^2 estimate of σ^2 is your MSE, so I can say now then this is my μ_L can I write down like this MSE/n. Then what I am saying that \bar{x}_L is the l-th population mean estimated from the sample. As it is an estimate, it is a random variable which follow a normal distribution with μ_L and MSE/n then I want to know the interval for μ_L .

So, you want to know the interval for μ_L , it is similar to finding out the interval of population mean what you have seen earlier using t test and other things can you write. What will be the this side and what will be the right hand and left hand side, can you tell me so that will be first will be the mean value point estimate value, what is required to know required to know, what type of distribution it is it is basically, we are saying that normal population and we will assume that sample size is small, then, what you require to do you require to use t-distribution.

Now what will be the degrees of freedom for this. In this case, you will find out that t-distribution degrees of freedom will be N-L, see and I think you can remember the earlier case when two population cases, we had gone for N_1+N_2-2 . So, here this N is N_1+N_2 upto N-L, so the same manner this $\alpha/2$, correct then what you require to write down this what is the L that is the variance part. So, σ^2 instead of σ^2 , I am writing MSE/n then this side, it will be $\bar{x}_L + t(\alpha/2)_{N-L} \sqrt{MSE/n}$.

So, this is what is the confidence interval for \bar{x}_l where x , sorry you please remember, it is always confidence interval related to the population parameter that is μ_l . So, it is a confidence interval for μ_l and the range is \bar{x}_l minus this quantity to \bar{x}_l plus this quantity, okay another issue what you want to do.

(Refer Slide Time: 09:30)

Handwritten notes on a blue background:

Random Variable $(\bar{x}_l - \bar{x}_m) \Rightarrow \mu_l - \mu_m$

$\leq \mu_l - \mu_m \leq$

$E(\bar{x}_l - \bar{x}_m) = \mu_l - \mu_m$

$V(\bar{x}_l - \bar{x}_m) = V(\bar{x}_l) + V(\bar{x}_m)$

$= \frac{\sigma_l^2}{n_l} + \frac{\sigma_m^2}{n_m} = \frac{\sigma^2}{n} + \frac{\sigma^2}{n} = \frac{2\sigma^2}{n}$

$= 2 \frac{MSE}{n}$

Logos for IIT KGP and NPTEL are visible in the corners of the slide.

You may be interested to know that what is the interval estimation for $\bar{x}_l - \bar{x}_m$ l-th population- m population that means what I mean to say, you will be interested to know $\mu_l - \mu_m$ and $\mu_l - \mu_m \leq$ to ≤ 1 this interval you want to find out and this one $\bar{x}_l - \bar{x}_m$, this is the random variable. Let me repeat that we are interested to know the interval estimation the difference between two population mean μ_l and μ_m , we will be using the statistics $\bar{x}_l - \bar{x}_m$.

So, what you require to know, you require to the expected value of $\bar{x}_l - \bar{x}_m$ this will be nothing but $\mu_l - \mu_m$ you also require to know the variance component of $\bar{x}_l - \bar{x}_m$, this will be variance of \bar{x}_l plus variance of \bar{x}_m and all of us know that this will be your σ I can write L here, $\frac{\sigma^2}{n_l} + \frac{\sigma_m^2}{n_m}$, you can write. But here what we have say consider that all the population means are equal and sample size is also same. So, we can write this one by $\frac{\sigma^2}{n} + \frac{\sigma^2}{n}$ that is $\frac{2\sigma^2}{n}$ which is nothing but $2MSE/n$.

(Refer Slide Time: 11:34)

$$(\bar{x}_l - \bar{x}_m) - t_{N-L}(\alpha/2) \sqrt{2MSE/n} \le \mu_l - \mu_m \le (\bar{x}_l - \bar{x}_m) + t_{N-L}(\alpha/2) \sqrt{2MSE/n}$$

$$(\bar{x}_l - \bar{x}_m) - t_{N-L}(\alpha/2) \sqrt{MSE(1/n_l + 1/n_m)} \le \mu_l - \mu_m \le (\bar{x}_l - \bar{x}_m) + t_{N-L}(\alpha/2) \sqrt{MSE(1/n_l + 1/n_m)}$$

100(1- α) % CI
Bonferroni Approach

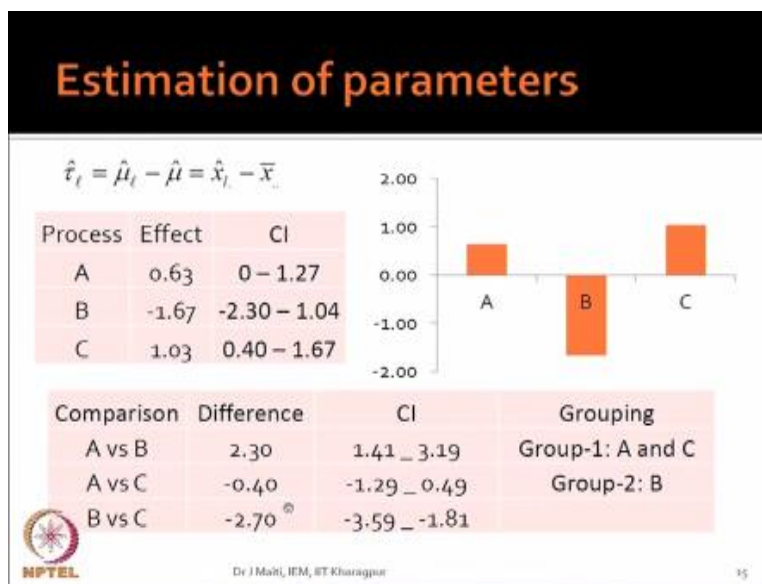
Now, if you want to find out now that what is my $\mu_l - \mu_m$ the interval then you must write down that this one, this is the first part then minus, what is the, we again consider the same t test. So, we will consider N-L $\alpha/2$ only thing change is instead of MSE/n, it will be 2MSE/n see the 2 is coming earlier for one mean case, you have to use MSE, 2 MSE/n. Now this side will be $\bar{x}_l - \bar{x}_m$, minus sorry plus $t_{N-L} \alpha/2 \sqrt{2MSE/n}$. What will happen if your sample sizes are different? This quantity will not be 2MSE/n exactly.

So, for unequal sample size case $(\bar{x}_l - \bar{x}_m) - t_{N-L}(\alpha/2) \sqrt{MSE(1/n)}$ suppose for the first one n_l , in l you write second one n_m because this l and m population less than equal to $\mu_l - \mu_m$ less than equal to $(\bar{x}_l - \bar{x}_m) + t_{N-L}(\alpha/2) \sqrt{MSE(1/n_l + 1/n_m)}$, okay. Now, see we by seeing this, what we are saying this 100(1- α)% CI, correct. Is it true that it is 100(1- α)% if I compare all the pairs, it will be reduced. So, it is basically we are talking about, okay two populations we are not com by saying in this we are talking about only two populations. We are not talking about that multiple comparisons simultaneously.

So what is required you are required to find out, how many comparisons possible and then you are doing go for Bonferroni approach. So, you can make it tighter by using Bonferroni approach,

okay. So, let us now again I will repeat one thing that basically first is collectively you are rejecting or accepting null hypothesis, where we are saying that all means are equal or one pair is different at least one pair is different. Then when you find out that null hypothesis is rejected you are finding out the inter p point as well as interval estimation for the population means as well as mean differences. Now, you require to know that if sum of x_0 is rejected then which of the pairs are different, correct.

(Refer Slide Time: 15:35)



So, we will see here with an example now, see the same example again we say that the process A B C the effect A effect is 0.6 then B effect is -1.67 and C effect is 1.03. If you sum total, this what is happening how much it is all effect sum will be 0, keep in mind this one here it is basically rounding error is there. So, 0.63, -1.67,+1.03 that is 0.01 or something is coming, but it is rounding error, okay. Then using the formulation, we have found out the confidence interval that is 100(1- α)%. Here, α is 0.5, see what is the value we are getting for A it can be 0 to 1.27 that means it is almost 0 it includes 0 that means no effect almost.

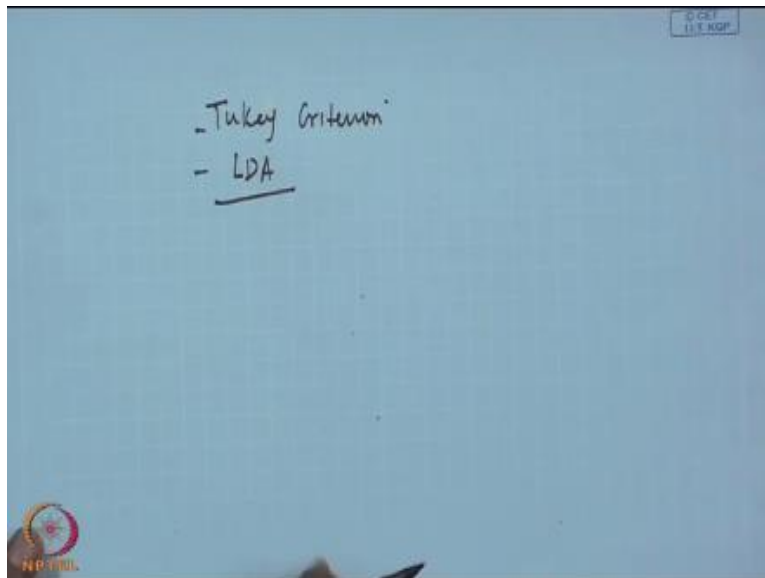
Second one you see it is -2.30 to 1.04, third one if you see it is 0.40 to 1.67 that is the confidence interval for the individual effects. Then we will find out that what is the difference in the

different population means okay, so here first is the process effect and confidence interval for mean, second one is the difference in population effect. The differences are for A to B that is 2.30 A to C -0.40 and B to C -2.70 what are these, these are the mean differences, correct.

Now, using this formula again using this formula, the second one as equal mean the second one this formula, we are computed the confidence interval. Now, in the first case A versus B the confidence difference in means the confidence interval is 1.41 to 3.19 in between there is no 0, see there is no 0 in between that means there is you can see that there is a difference between A and B. Now, come A to C you see A to C -1.29 to 0.49 in between there is 0, so A C difference is not there or less from strategical sense we will say it contains 0 means there is no difference.

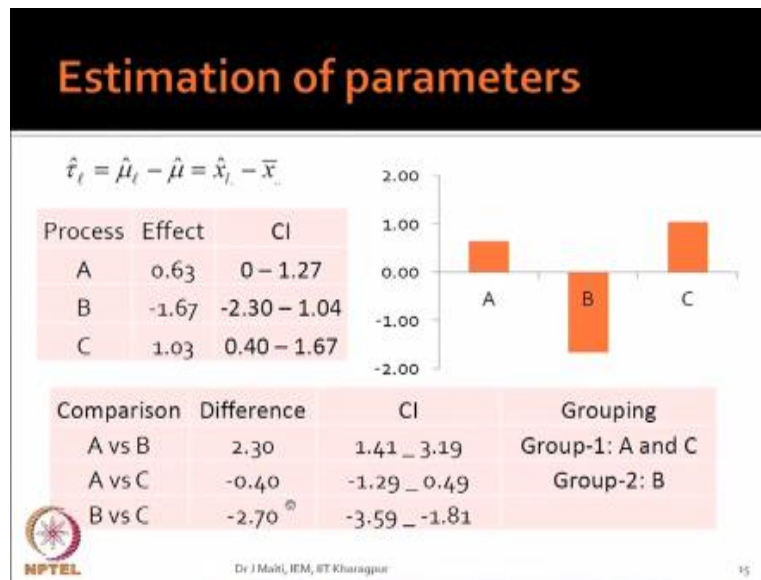
What about B to C minus to minus, so it is there. So now if we group we can say that A and C they are almost equal effect, but B is entirely different that is what you want to understand, okay. So, if you use Bonferroni approach, it will be a little tighter, but there is another many other techniques available.

(Refer Slide Time: 19:25)



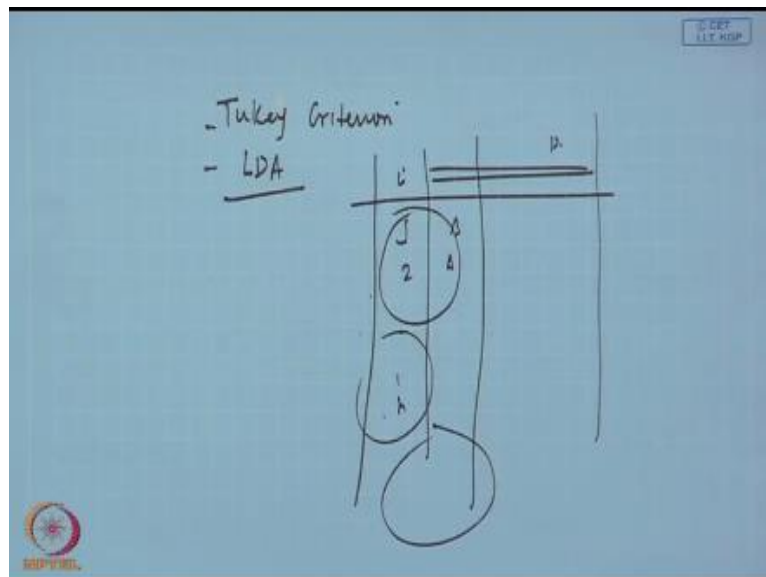
Like Tukey criteria, I think you know Fischer LDA linear discriminant analysis, which one this concept, which one you are talking about that grouping, that one and A and C all those things.

(Refer Slide Time: 19:53)



Clustering see ultimately you are some groups you are getting essentially clustering means from the data you want to group something. So, you want to find out I have different items I want to group, but here the primary difference is not that you are process A and process B and process C, you are talking about one observation from each procedure, this is not you are taking a large number of values in term of samples from the processes. So essentially, you are not grouping the items in that sense.

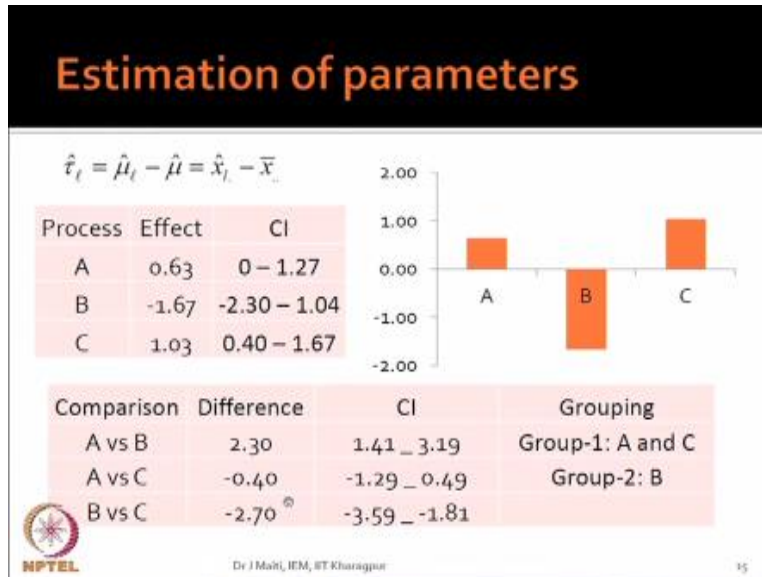
(Refer Slide Time: 20:44)



Clustering is what clustering is suppose, I have several individuals 1 to n then because depending on certain characteristics. Let it be p characteristics are there, it may show up in that 1 to 5 these are making one group rest is one group rest is one group, okay that is the from individual observation point of view we will find out like this here. So, here also you can say because what is happening here instead of i , if I say the first one is A, second one is A like this, basically in three different groups are there I think it is not fit for clustering.

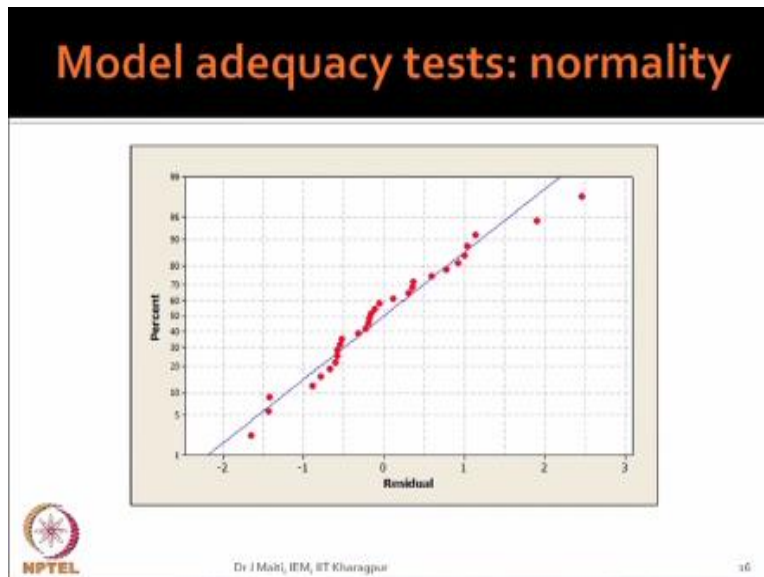
The reason is we know A B C clustering essentially what would clustering you are saying supervised or unsupervised one, unsupervised case groups are not known. But here we know that this is groups are there A B C are there, but again with A B C you are grouping further because clustering is very rich. So, in clustering time you use either has radical clustering or k means clustering C mean clustering that are there superior method.

(Refer Slide Time: 22:05)



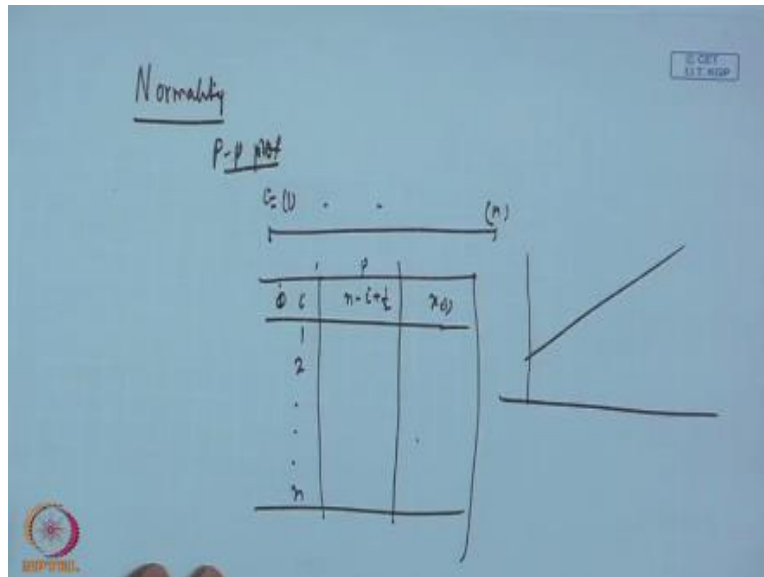
We should not go in that this manner, but again definitely I can say you have to inform of here two groups and that is why the third in this line that can it be clustered that is the grouping is basically here we are saying from mean point of view they are similar, okay.

(Refer Slide Time: 22:30)



Then, you require seeing that whether your model is adequate or not, how do you understand my model is adequate, first of all you have to test the assumptions.

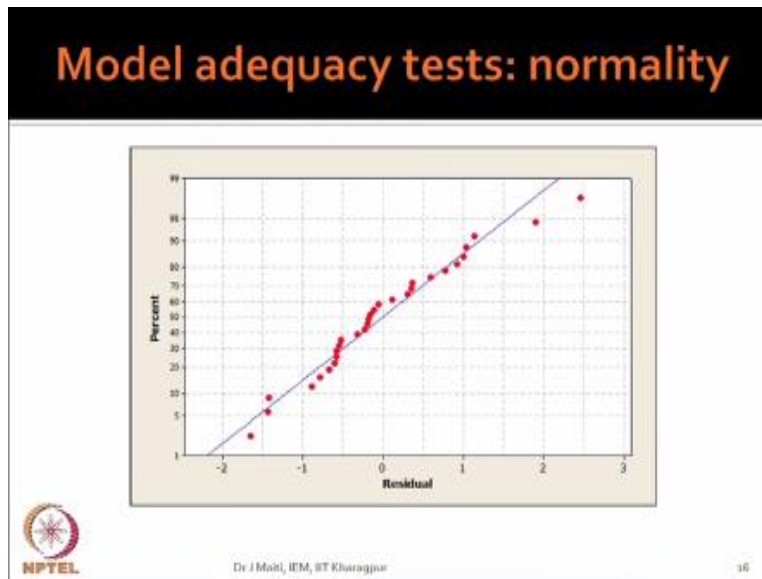
(Refer Slide Time: 22:59)



So, under ANOVA one of the assumptions is normality. How do you test normality, any idea? Quantile plot is there, that is for multi variate case, univariate case you can go for quantile, but that will be z quantile. Otherwise, simple p plot probability, probability plot is there, here what will happen in probability, probability plot, I think you go through any basic statistics book. You will be finding out, you will basically compare the empirical probability with the observed values, first you arrange the values from lowest to highest.

Suppose, your i equal to that 1 to n , let they are already ordered then what will happen you will find probability that i , I think you will be look like this $\frac{i-1}{n}$, $\frac{i+1}{n}$, $i=1, 2$ to n and this is the empirical probability, cumulative probability basically there is a cumulative probability then you have already that values are there suppose x values are there or x ordered values are there, so you plot this with this and you must get a straight line, this is probability, probability plot here what is happening in this figure, and you see.

(Refer Slide Time: 24:43)



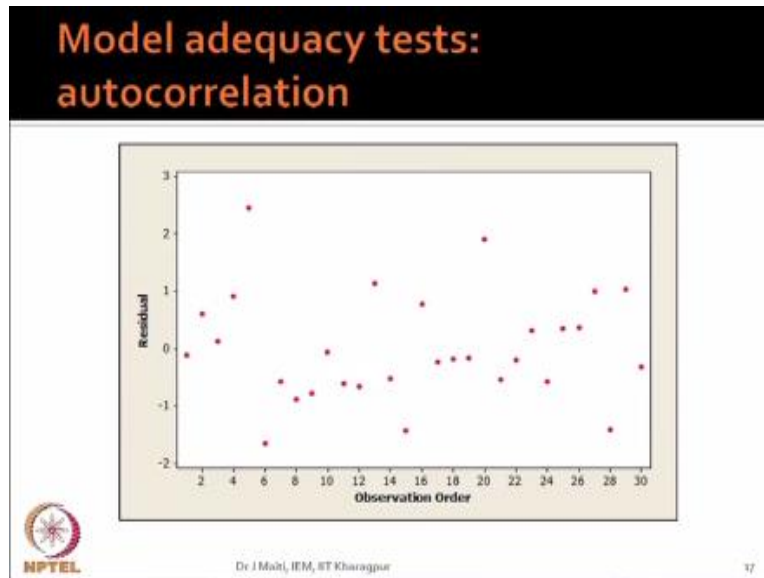
Histogram what will happen you will not get that the from the histogram, you get a feel that, okay that may be normally distributed, but you cannot say 100% that it is normal or not. Similarly probability plot is better than histogram here, but again probability plot also it is very difficult to know that what is the extend of departure. So, in that case you have to go for K-S test Kolmogorov-Smirnov test. Now you see this probability plot given here. This is the residual see residuals are arranged from lowest to highest this is the ordered one and this side the probability that is basically cumulative part percentage.

So, if really your n is 10 then the first one will be that is in the fifth percentile case rest will be 95 will be rest that thing. So, in that sense the far few fifth, tenth then 15 if number of observation is 10. So, in that way you are plotting here, so you will get a straight line if you do not get a straight line then it says that there is departure from normality. For example, in this case last two values are this value and this value plus first value, but little departure I think that this is not gross departure.

So, we can assume that they are normally distributed, here again the subjectivity is there that we are saying, it is normally distributed but quantitatively. I think you can use z test is there

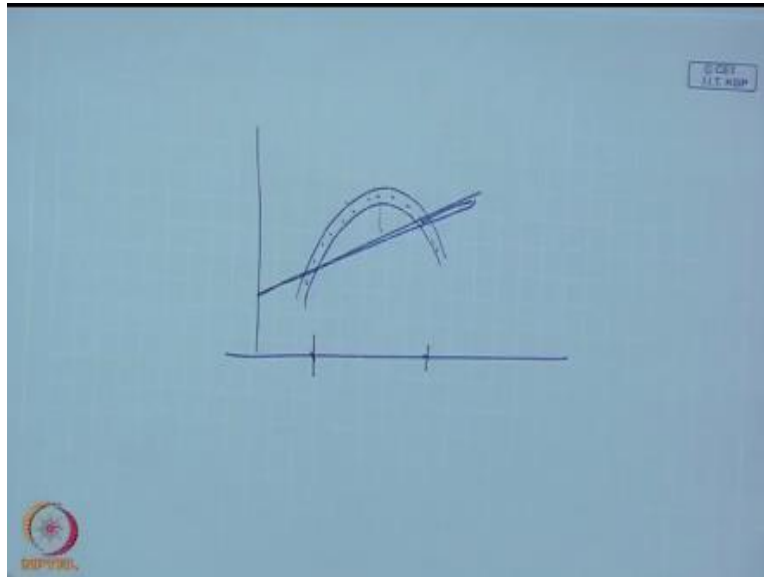
then your KS testis there all those things you can do, so for the data set normality test says that, yes, data comes from normal distribution.

(Refer Slide Time: 26:52)



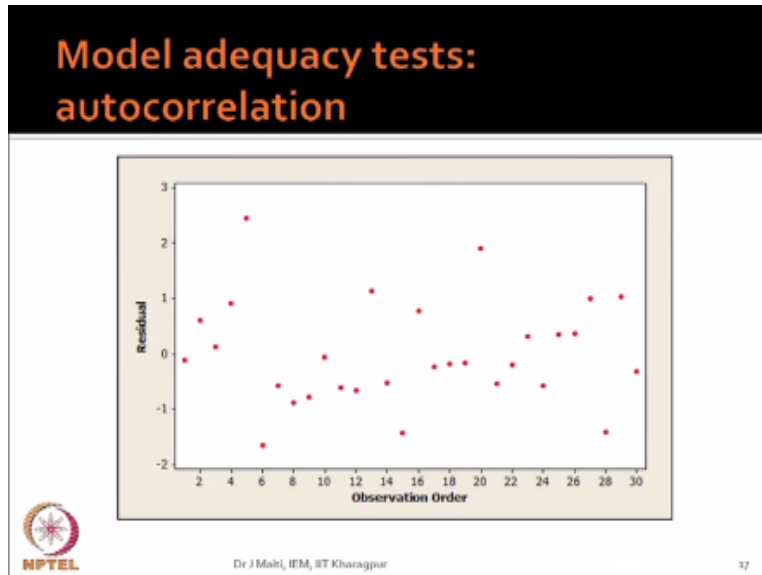
Then, auto correlation what is say observation should not be serially correlated. For example, first observation is related with the 1-th observation or i-th observation that should not be correlated it may happen that overtime li you collect something there will be time lag may be January data to next month, January data, next year January data that will be that may be seasonal correlate correlation. So, that type of case what will happen when you fit a linear model like ANOVA, then the entire non linearity part will go to the errors.

(Refer Slide Time: 27:40)



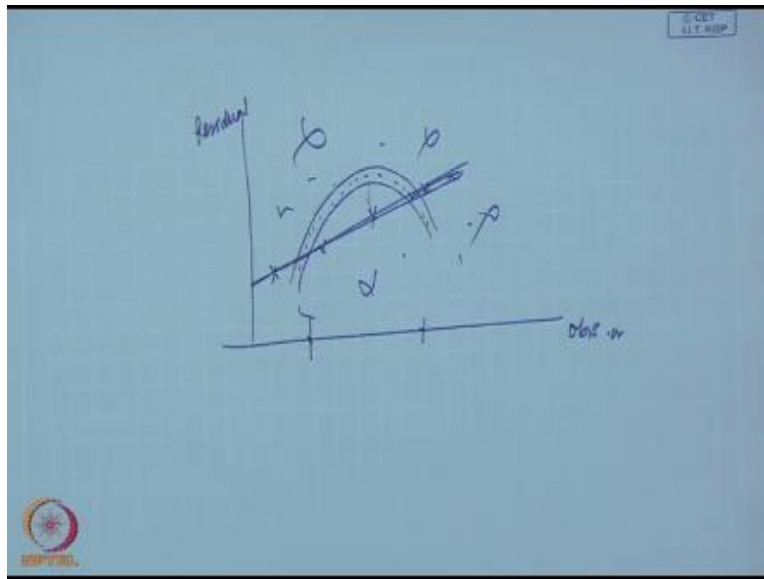
For example, you are fitting a linear model like this my data is like this, so ultimately you will capture this much, but this nonlinear depart where it will go to the aero. So, in this aero dome that will be seen, and another issue when I talk about that autocorrelation part when you talk about if there is correlation between this observations. This observation what will happen, this correlation part will not be captured here, but it will go to the aero rule capture all those things which are not captured by the model.

(Refer Slide Time: 28:29)



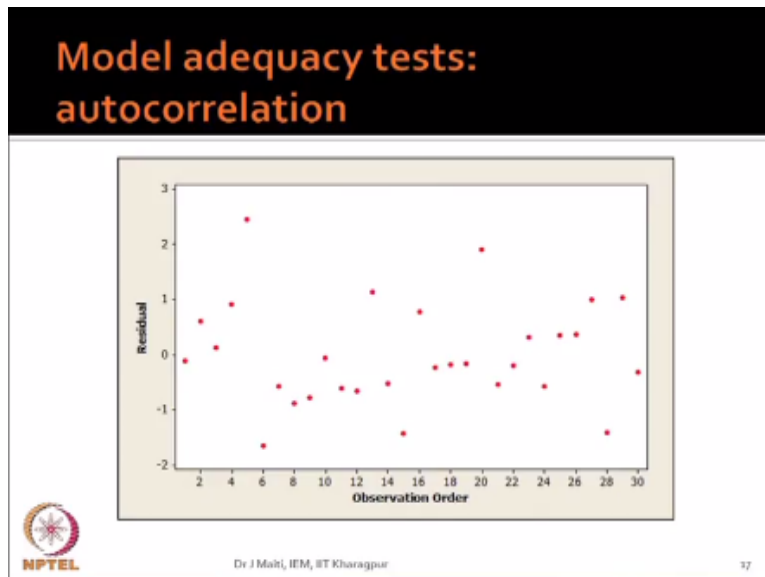
So, if I see the zero observation order which is these the residuals there should not be any systematic pattern. In this case is there any pattern are you able to see that here is a linear or nonlinear or some increasing or decreasing trend. I do not think, it is haphazard random one these are the observation in terms of your collection observation first, observation second, observation third like this observation order. If there is correlation then what will happen.

(Refer Slide Time: 29:14)



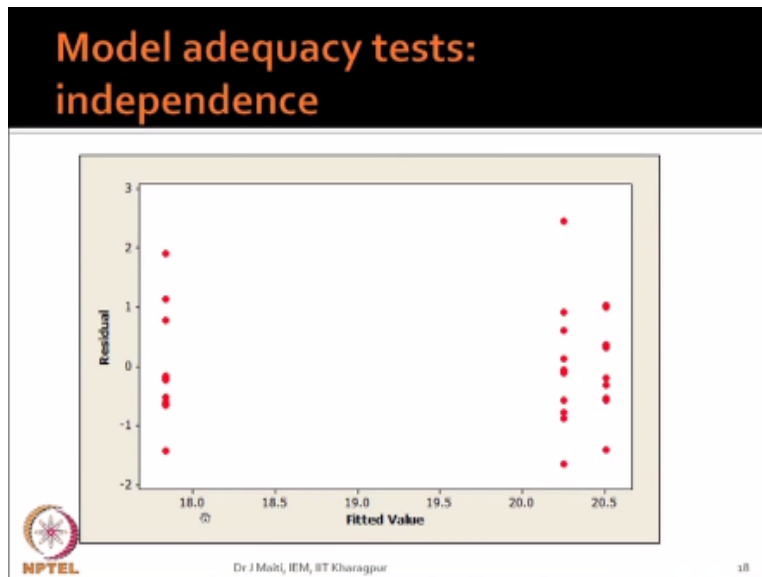
Suppose if this side is residual and this side is your observation order. So, if your values are like this what is happening then they are correlated even if values will come like this they are correlated, but nonlinear relation is this should not happen. It should be something like this nobody knows where it will be.

(Refer Slide Time: 29:42)



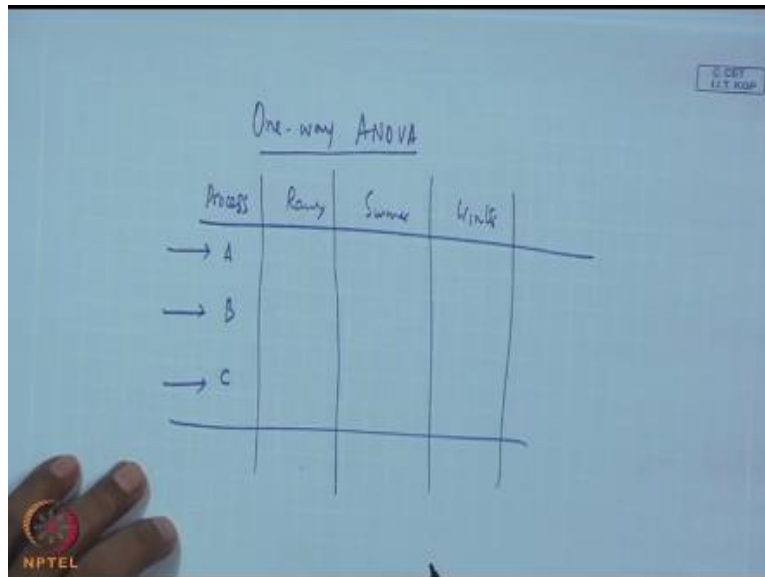
So this says that our data is definitely independent.

(Refer Slide Time: 29:50)



Then I think this is autocorrelation, then we go for independent test, in independent test you are making the residual versus fitted value plot, okay. Actually, what happened in this case same values for a particular fitted value, this is a fitted, this is where you will be getting like this because here it is our data is such that you will be getting in a particular point there are so many residuals values are there, but if you take large number of data you will be finding out again no values, it will be filled up all the plot total plot will be filled up by the data points, so there is no pattern again, so no problem then they are independent.

(Refer Slide Time: 30:45)

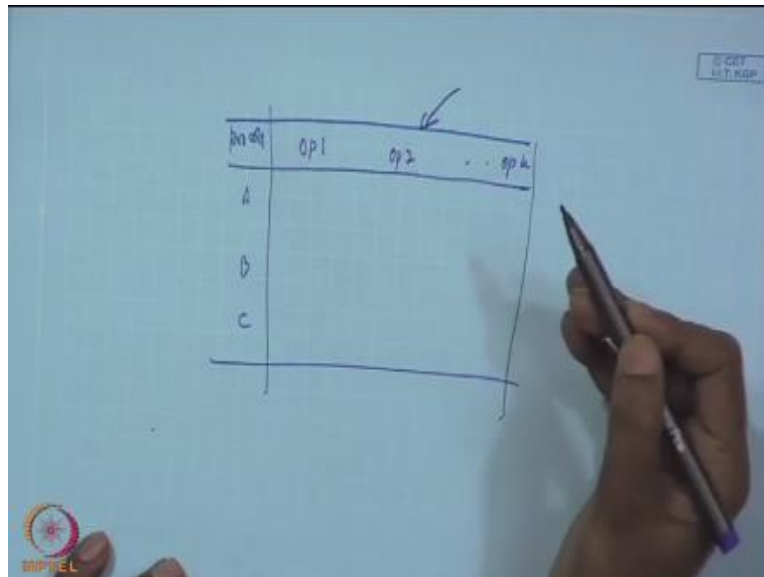


The image shows a hand-drawn table on a whiteboard titled "One-way ANOVA". The table has three columns labeled "Process", "Rainy", "Summer", and "Winter". The rows are labeled "A", "B", and "C". The "Process" column is empty, and the "Rainy", "Summer", and "Winter" columns are also empty. A hand is visible in the bottom left corner, and an NPTEL logo is in the bottom left corner of the whiteboard. A small box in the top right corner of the whiteboard contains the text "C. GUY" and "11.11.2018".

Process	Rainy	Summer	Winter
→ A			
→ B			
→ C			

So, far we have discussed ANOVA which is known as one way ANOVA, getting me one way ANOVA, why one way because we have considered only different population in terms of one factor. For example, I say that different processes we say that process A, process B, process C like this and we treated them as if this is population 1, population 2 and population 3 now what is the guarantee that the process will perform equally over the seasons. It may so happen that during rainy season, the performance will be bad than summer and winter this is one.

(Refer Slide Time: 32:03)



A hand-drawn table on a whiteboard. The table has a header row with columns labeled 'Inst', 'Op 1', 'Op 2', and '... Op k'. Below the header, there are three rows labeled 'A', 'B', and 'C'. A hand holding a pen is visible on the right side of the table, pointing towards the 'Op 2' column. In the top right corner of the whiteboard, there is a small rectangular stamp that reads '© CRYSTAL' and '© T. K. S. P.'. In the bottom left corner, there is a circular logo with a star and the text 'MPPCL'.

Inst	Op 1	Op 2	... Op k
A			
B			
C			

Second thing is that, second issue maybe suppose the process operators process A, process B, process C, ABC is there, but there will be different level operator level 1, operator level 2 and operator level. Suppose your k I mean who is actually operating the process that also determines the quality of the process. So, in that case, the first case and the second case what you required to do, you require to include the effect of operator maybe the effect of the seasons.

(Refer Slide Time: 32:43)

One-way ANOVA Noise Variable

	Rainy	Summer	Winter
→ A			
→ B			
→ C			

NPTEL

Now, see you do not have control on the seasons, so even though you know, but you maybe some action plan you may be taking, but one over that seasons you have no control. So, in that sense what will happen sometimes, suppose we can say that season is a noise variable, getting me. But you want to estimate the process effects keeping in view that the seasonal effects to be blocked, getting me what I am trying to say.

So, that means you will one things is that you go for any season production and see the quality summer production see the quality, winter production season see the quality that means, you do not want the noise variable effect to be accumulated in the process effects. If you do like this type of design is known as blocking, so this is one variable is blocked variable is there, so rainy seasonal different, summer is different and winter is different.

(Refer Slide Time: 34:02)

Process	Op1	Op2	Op3
A			
B			
C			

It may so happen that your control over the operator so depending on the situation, you may change the operator you give the operator to different machines so and may be fit, so in that case if you have control over the operator. So, this process is one factor which is controlling the quality of the product produced and as well as operator is another factor that is factor 2.

(Refer Slide Time: 34:35)

One-way ANOVA

	Process	Rawy	Summe	Noise Variable Factor 2
Factor	→ A			Levels
	→ B			
	→ C			
				Blocks

Even here, we can say noise variable this is factor 2 and process is factor 1.

(Refer Slide Time: 34:45)

A hand-drawn Gantt chart on a whiteboard. The chart has three rows labeled 'A', 'B', and 'C' and four columns labeled 'Activity', 'OP1', 'OP2', and 'OP3'. The 'Activity' column contains 'A', 'B', and 'C'. The 'OP1' column contains a horizontal bar for activity A. The 'OP2' column contains a horizontal bar for activity B. The 'OP3' column contains a horizontal bar for activity C. There are arrows pointing from the word 'Factor' above the chart to the 'OP1' and 'OP2' columns. A hand is pointing at the chart from the right side.

Activity	OP1	OP2	OP3
A	Bar		
B		Bar	
C			Bar

Only difference between the first and second is blocking in this.

(Refer Slide Time: 34:48)

One-way ANOVA

	Process	Raw	Season	Units
Factor → A				
→ B				
→ C				

Handwritten notes: "Main Variable Factor 2" with an arrow pointing to the "Season" column, and "Block" written at the bottom right.

In this case you are not interested to know or you are not in a position to know that this one, better is you are not interested to know the seasonal effect, because you do not have control on seasons. So, nullifying the seasonal effect or eliminating the seasonal effect, you want to estimate the process effect.

(Refer Slide Time: 35:13)

The image shows a handwritten diagram on a whiteboard. At the top, a table is drawn with columns labeled 'Op1', 'Op2', and 'Opk'. To the left of the table, the word 'Factor1' is written vertically. The rows of the table are labeled 'A', 'B', and 'C'. To the right of the table, the text 'Two-way ANOVA' is written. Below the table, the following text is written: 'Factor 1 - a levels, 1, 2, ..., a.' and 'Factor 2 - b levels, 1, 2, ..., b.'. In the bottom left corner, there is a small circular logo with the text 'NPTEL' below it.

	Op1	Op2	Opk
A			
B			
C			

Factor 1 - a levels, 1, 2, ..., a.
Factor 2 - b levels, 1, 2, ..., b.

Here you want to estimate both seasonal effect, sorry operators effect plus process effect. This type of design is two ways ANOVA, this is two way ANOVA, what will happen here, why two way because factor 1 having suppose a levels that means factor 1 is having 1, 2, 3 like up to a levels, factor 2 is having b levels. That means 1,2 like up to b levels, then your design is like this.

(Refer Slide Time: 36:00)

Factor 1	Factor 2				$\bar{x}_{L.}$	$\tau_{L.}$
	1	2	...	b		
1	x_{11}	x_{12}	...	x_{1b}	$\bar{x}_{1.}$	$\bar{x}_{1.} - \bar{x}$
2	x_{21}	x_{22}	...	x_{2b}	$\bar{x}_{2.}$	$\bar{x}_{2.} - \bar{x}$
...
a	x_{a1}	x_{a2}	...	x_{ab}	$\bar{x}_{a.}$	$\bar{x}_{a.} - \bar{x}$
	$\bar{x}_{.1}$	$\bar{x}_{.2}$...	$\bar{x}_{.b}$	\bar{x}	
	$\bar{x}_{.1} - \bar{x}$	$\bar{x}_{.2} - \bar{x}$...	$\bar{x}_{.b} - \bar{x}$		

What is your design, in this side, you are writing factor a, this side or factor 1, this side factor 2, factor a or factor 1 and factor 2, so factor 2 is having b levels, factor a is having a levels or factor 1 is having a levels, okay. Usually, we denoted factor a or factor 1, factor b, ab is coming like this, now what will happen you will collect data, okay. So, factor 1 at level 1, factor 2 at level 2 you collect a large number of data, correct here also you collect some data.

So, everywhere you are collecting data, if I go by the population concept that what we have discussed in ANOVA, then what you are getting here you are getting two important things, one is for factor 1 you will be getting here x , if I say this is \bar{x}_L , I am giving one dot here, what is this I, you can ignore factor 2, if you ignore factor 2 then the sum total of observations here this L, this L, this L, this L, this observation is related to 1. So, you will be getting 1 average, you are ignoring that to 2.

Similarly here you are getting 2 averages like this here you will be getting a average, now you ignore a, you talk about factor 2 only, then here also you will be getting average. So, this one we are denoting as $\bar{x}_{.m}$, $\bar{x}_{.1}$ stands for all the rows here that are levels for factor 1. $\bar{x}_{.m}$, m stands for levels in factor 2. So, then this will be $\bar{x}_{.1}$, $\bar{x}_{.2}$, so like this I think this one $\bar{x}_{.b}$, correct. Then here

will one grand average, yes or no. Grand average means, I do not consider a or b anything or factor a 1 or factor 2. I will consider total these are all my observations, so it will be basically average of all the values given here.

Now as you have computed this \bar{x}_1 here ignoring this factor 2 so, you can find out that τ_1 . this minus grand, similarly here also we can find out yes or no. I think we will not use this one dot, we will not use here τ_1 and similarly here we will use τ_m , then this one will be your $\bar{x}_{.1} - \bar{x}$, $\bar{x}_{.2} - \bar{x}$ like this $\bar{x}_{.b} - \bar{x}$ this is known as two ways ANOVA, getting me two way one factor1, 1 a levels factor b, factor 2 b levels and there is another issue is known as interaction between the factors. Now, what will be your ANOVA table in this case? In this case your ANOVA table will we like this.

(Refer Slide Time: 41:01)

Source of variation	SS	DF
Factor 1	SS1	a-1
Factor 2	SS2	b-1
Interaction (12)	SS12	(a-1)(b-1)
Error	SSE	
Total	SST	abn-1

Source of variation, one is factor 1, factor2 is also source, interaction 1 and 2 between factor 1 and factor 2 and error then total, here you have to find out sum square, so it is I can write here sum square a I am writing for the first factor, second one I am writing sum square b. I think we should change little different because b already before between we have taken b. So, factor 1, factor 2, factor2, sorry if I write like this SS1 and SS2 no problem, then factor 2 then SS12, then

it is SSE and this will be SST. ANOVA case, it is always basically partitioning the observations into different components, similarly partitioning the total variability in to the sources variability, okay. So, here what will be the degrees of freedom, how many levels are there for factor 1 a level, so a-1, how many levels are there for factor 2, b-1 what will be the interaction, a-1, b-1, what is the total number of observation abn-1.

(Refer Slide Time: 43:00)

$c = a-1 \quad b = b-1 \quad df$

Factor 1	Factor 2				
	1	2	...	b	$\bar{y}_{.j}$
1	\bar{y}_{1j}
2	\bar{y}_{2j}
...	$\bar{y}_{.j}$
a	\bar{y}_{aj}
					$\bar{y}_{..}$

Because see if we consider that you are collecting n observation, here n observation here, everywhere n observation. There are levels for factor 1, b levels for factor 2, there are ab cells and in each cells there are n observations, so abn.

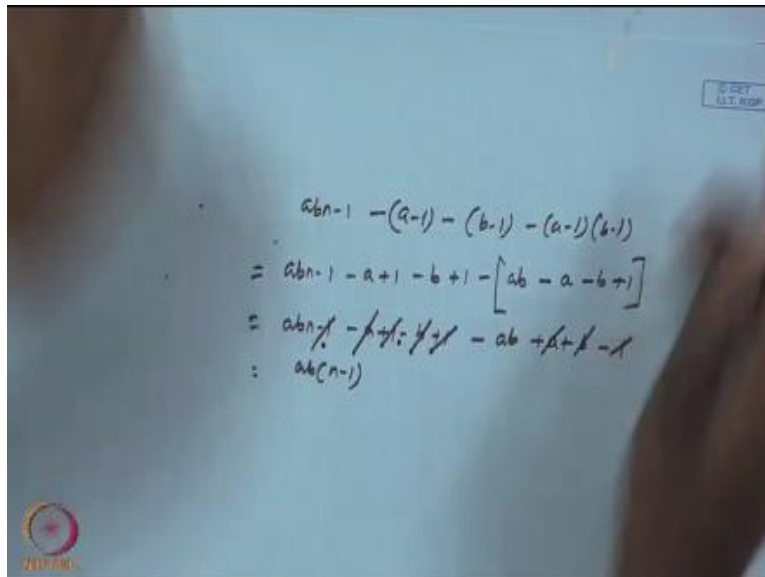
(Refer Slide Time: 43:19)

Source of Variation	SS	DF
Factor 1	SS_1	$a-1$
Factor 2	SS_2	$b-1$
Interaction (12)	SS_{12}	$(a-1)(b-1)$
Error	SSE	
SST		$abn-1$

N = abn

So that is what is our N, N is abn, now if you collect different samples for different shells and as well as if you collect different samples for the factor1 and factor 2 differently, ultimately this computation will be little different but the, you can still write n, you have to compute n, okay then what will be the SSC space $abn-1$, $-(a-1)$, $-(b-1)$, $-(a-1)(b-1)$ because total will be this. So, what will be this value.

(Refer Slide Time: 44:02)



The image shows a whiteboard with handwritten mathematical work. The work consists of four lines of equations:

$$\begin{aligned} & ab^{n-1} - (a-1) - (b-1) - (a-1)(b-1) \\ &= ab^{n-1} - a + 1 - b + 1 - [ab - a - b + 1] \\ &= ab^{n-1} - \cancel{a} + \cancel{1} - \cancel{b} + \cancel{1} - ab + \cancel{a} + \cancel{b} - \cancel{1} \\ &= ab^{n-1} \end{aligned}$$

In the top right corner of the whiteboard, there is a small rectangular stamp that reads "SECRET U.S. EYE". In the bottom left corner, there is a circular logo with a red and blue design and some text below it.

$ab^{n-1} - (a-1) - (b-1) - (a-1)(b-1)$, okay, so this $ab^{n-1} - a + 1 - b + 1$ so minus this is ab , okay $-a - b + 1$, correct. So, I am writing $ab^{n-1} - a + 1 - b + 1 - ab + a + b - 1$, so $+a - a$, $+b - b$, $-1 + 1$ then it is basically this is $1 - 1$ know, $+1$, this also cancelled out, so that mean ab^{n-1} .

(Refer Slide Time: 45:07)

Source of Variation	SS	DF	MS	F	Decision
Factor 1	SS1	a-1	$MS1 = \frac{SS1}{a-1}$	$F_1 = \frac{MS1}{MSE} > F_{a-1, ab(n-1)}$	Reject H_0
Factor 2	SS2	b-1	$MS2 = \frac{SS2}{b-1}$	$F_2 = \frac{MS2}{MSE} > F_{b-1, ab(n-1)}$	- Do -
Interaction (1x2)	SS12	(a-1)(b-1)	$MS12 = \frac{SS12}{(a-1)(b-1)}$	$F_{12} = \frac{MS12}{MSE} > F_{(a-1)(b-1), ab(n-1)}$	- Do -
Error	SSE	ab(n-1)	$MSE = \frac{SSE}{ab(n-1)}$		
Total	SST	abn-1			

N = abn

So your, okay then what you require to know you require to compute MS mean square that variability this will be MS1 will be $SS1/a-1$, MS2 will be $SS2/b-1$ then MS12 that is interaction which is $SS12/(a-1)(b-1)$ then your MSE will be $SSE/ab(n-1)$ then you find out F, what will be your F, F basically the concept of F here is here why we will use in this fashion F, what do you want to see that whether factor 1 variability explanation is more than the error, or not. Whether factor 2 is explaining equally to error or it is better or interaction to error.

So, that means every sum squares here, we will take the mean squares, we will be compared with the error, so as a result for these F1 if I say for the factor 1, this one will be $MS1/MSE$ for factor 2 it will be $MS2/MSE$ for 1, 2 it is $MS12/MSE$. These are the things you want to test, okay. Now what will be the degrees of freedom for the this case, if this one is greater than what is our numerator degrees of freedom a-1 and degree of freedom is $ab(n-1)$. If this α this is the case, so if computed F is greater than the tabulated F then what will be your decision reject H_0 .

So, similarly if this $F_{b-1, abn-1}$ do it, if this one $F_{a-1, b-1}$ I think into $(a-1)(b-1)$ and $abn-1$, then do this is what is known as two way ANOVA. Now, you can go for multi way ANOVA also what

will happen ultimately, suppose a three way ANOVA what you will do how to partition it three way ANOVA case,

(Refer Slide Time: 48:30)

Factor C -

Source of variation		SS	DOF	F
→ A	Main effects	SSA	a-1	F
→ B		SSB	b-1	
→ C		SSC	c-1	
→ AB	2-way interaction effects	SSAB	(a-1)(b-1)	
→ AC		SSAC	(a-1)(c-1)	
→ BC		SSBC	(b-1)(c-1)	
→ ABC	3-way interaction effects	SSABC	(a-1)(b-1)(c-1)	
→ Error		SSE		

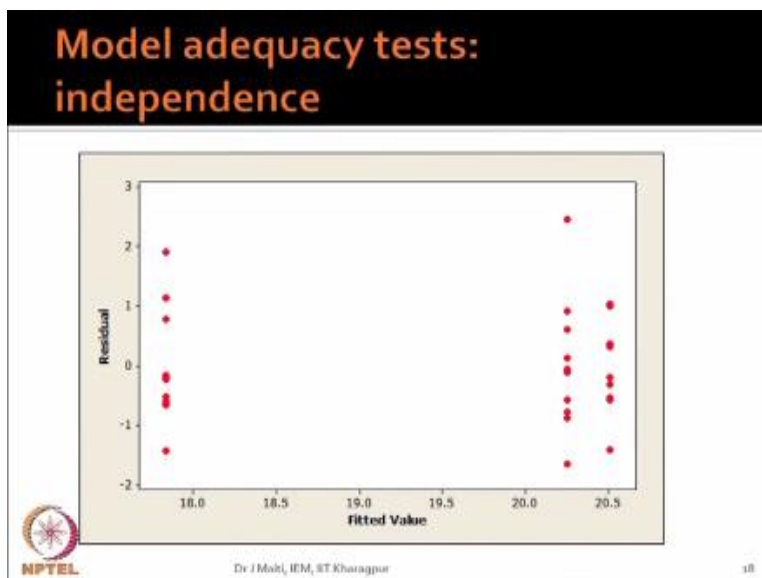
Three way ANOVA, that mean if I say factor I am now writing in terms of ABC, factor A with a levels, factor B with b levels, factor C with c levels, so if I collect equal sample size suppose sample size is n, so your total observation will be abcn, total observation that you get, okay. Now, which I will not we will not compute anything; only thing I want to write here what will be the sources of variation, sources of variation. Definitely first is A, second is B, third is C, then definitely AB interaction between A and B, interaction between A and C, interaction between B and C interaction ABC, correct.

See, first is A, second is B, third is C, AB may interact AC may interact BC may interact and all three may interact class error. So, your total variability will be decomposed into different sources, now ABC, these are all known as main effects AB, AC, BC in this case will be two way interaction effects. ABC is three way interaction effects and error, so you have to SS for everywhere you have to find SS then you have to find DOF, degree of freedom. Degree of freedom is for main effect a-1, b-1, c-1, interaction of it (a-1)(b-1) for AB, for AC (a-1)(c-1) like

this, ABC case (a-1)(b-1) that multiplication and your toe for your total case that will be $abcn-1$. Then the for error part just subtract once you get all these things, suppose SSA, SSB, SSC, SSAB, SSBC, SSAC, SSE, SST with all degrees of freedom you have to compute it F, F_a, F_b like this and you know what will be the degrees of freedom for the tabulated value that F value, we calculate this is what is the decomposition and by this manner you will be able to find out the effect of many factors the interactions.

It may so happen that your case is such that there is no effect for B or C, but A, AB, BC these all bits are there, so that interaction effects you have to control, getting me any question you want to ask me. No question, I think now you are you are in a position to explain ANOVA for any number of factors just you compute little bit. For three digit, 2^k , suppose I have that is it because it all it all depends because three factors and how many main effects three main effects, how many interaction effect, so that mean and see 1 is coming, then two at a time, and see 2 is coming, okay, that we will work it. So, that is three factors three at a time, three factors three at a time, will be one at a time will be three and two at a time will be another three. So, that is in set is coming, so you calculate this, okay.

(Refer Slide Time: 53:57)



Ask me some question. So, this is basically the totality from ANOVA point of view, I think ANOVA is very, very useful, very, very powerful technique also, and you find out in many cases ANOVA is required, but this is a univariate model keep in mind. In the multivariate counterpart is MANOVA.

(Refer Slide Time: 54:25)



So, we will discuss in next class MANOVA will be discussed in next class.

(Refer Slide Time: 54:36)



References

- D C Montgomery, Design and Analysis of Experiments, Wiley India Edition, New Delhi, 2012.
- Johnson R A and Wichern D W, Applied Multivariate Statistical Analysis, PHI Learning Pvt. Ltd., New Delhi, 2013.

 Dr J Mohi, IEM, IIT Kharagpur 39

And for ANOVA, I suggest you to follow DC Montgomery design and analysis of experiment. This is the best book I have read in terms of your ANOVA, large number of real life examples are given and if you know ANOVA, MANOVA will be extension to the multivariate domain little bit computational difficulty will always be there, because you will be computing in the matrix domain. This all sum square will become, that also we have seen the matrix domain sum square will not be there, it will be SSCP sum squares and cross product, okay. 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