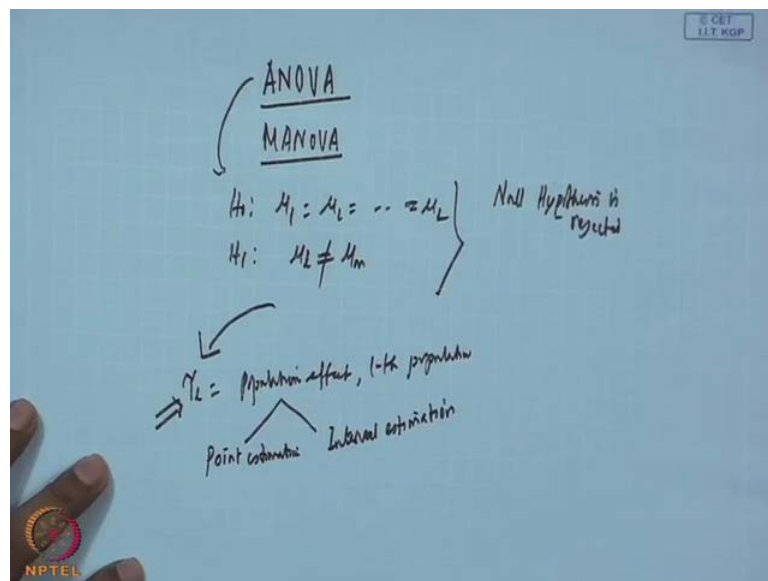


Applied Multivariate Statistical Modeling
Prof. J. Maiti
Department of Industrial Engineering and Management
Indian Institute of Technology, Kharagpur

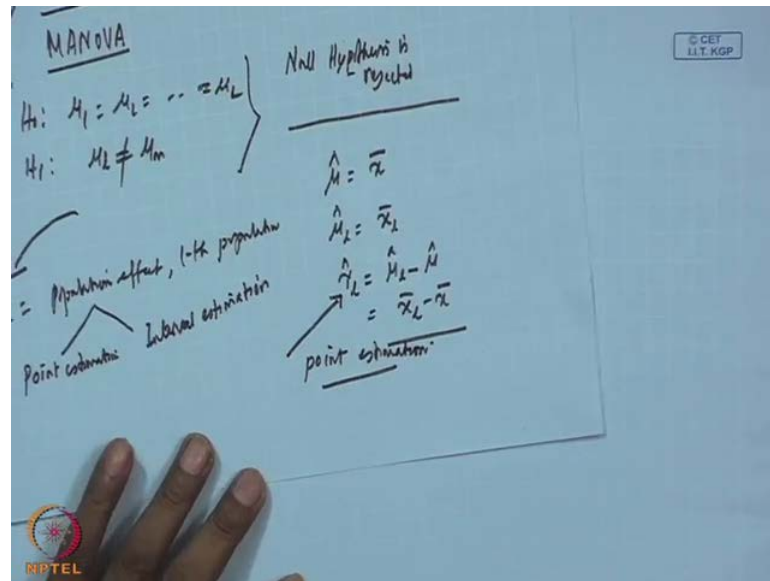
Lecture - 15
Analysis of Variance (ANOVA)
(Contd.)

(Refer Slide Time: 00:22)



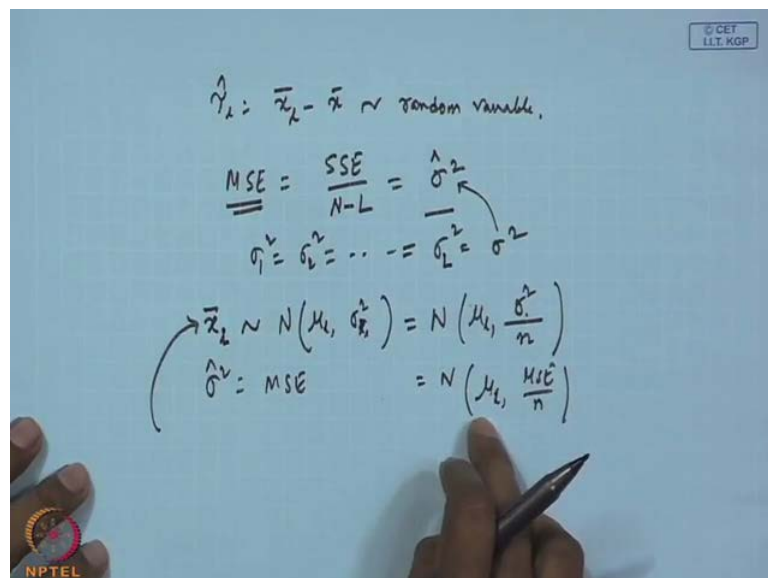
We will continue ANOVA for another 15, 20 minutes then we will discuss MANOVA. We have seen the hypothesis testing using ANOVA, where we have said that all the means are equal and alternative hypothesis at least one pair of means are not equal, and using F test, we either accept or reject the null hypothesis. For the problem, consider we find that the null hypothesis is rejected, so in our case the null hypothesis is rejected, hypothesis is rejected. Now, we will discuss two things; one is that when you talk about tau 1, that is the population effect that is the population effect that is called 1 th population, 1 population. We want to compute it how to compute this tau 1, 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.

(Refer Slide Time: 02:20)



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. Estimate of population effect is your estimate of pop that is the population mean minus grand mean, which is \bar{x}_i minus \bar{x} . So, this I can say that point estimation, this is 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)



We can say that the tau l which is your $\bar{x}_l - \mu_l$, this is a random variable, 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, tau l cap or mu l cap 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 minus l. This is estimate of sigma square, what is this sigma square. You have recall the bullet test where we said that sigma 1 square equal to sigma 2 square, all population variances are equal, equal to sigma square. So, this one is the estimate of sigma square, so MSE.

We know MSE is known, now if I want to know what the distribution of \bar{x}_l is. Then we will say this will follow normal distribution with mu l and that sigma l that sigma \bar{x}_l square this one and you can remember that this will be nothing but mu l into sigma square by your n. Now, sigma square estimate of sigma square is your MSE, so I can say now then this is my mu l can I write down like this MSE by 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 that is follows normal distribution with mu l and MSE by n then I want to know the interval for mu l.

(Refer Slide Time: 06:44)

Handwritten notes on a whiteboard:

- Top line: $T_l = \bar{x}_l - \mu_l \sim \text{Random variable}$
- Second line:
$$\underline{\underline{MSE}} = \frac{SSE}{N-L} = \frac{\hat{\sigma}^2}{n}$$
- Third line:
$$\sigma_1^2 = \sigma_2^2 = \dots = \sigma_L^2 = \sigma^2$$
- Fourth line:
$$\bar{x}_l \sim N(\mu_l, \sigma_l^2) = N\left(\mu_l, \frac{\hat{\sigma}_l^2}{n}\right)$$
- Fifth line:
$$\hat{\sigma}_l^2 = MSE \quad = N\left(\mu_l, \frac{MSE}{n}\right) \leftarrow$$
- Sixth line:
$$\Rightarrow \underline{\bar{x}_l - t_{N-L}(\alpha/2) \sqrt{\frac{MSE}{n}}} < \mu_l < \bar{x}_l + t_{N-L}(\alpha/2) \sqrt{\frac{MSE}{n}}$$

The whiteboard also features an NPTEL logo in the bottom left corner.

So, you want to know the interval for mu 1, 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 minus 1 see and I think you can remember the earlier case when two population cases, we had gone for n 1 plus n 2 minus 2. So, here this capital n is n 1 plus n 2 up to n 1 minus 1, so the same manner this alpha by 2 correct then what you require to write down this what is a con that is the variance part.

So, sigma square by n instead of sigma square, I am writing MSE by n then this side, it will be x 1 bar plus t alpha by 2 n minus 1 square root of MSE by n. So, this is what is the confidence interval for x 1 bar where x, sorry you please remember, it is always confidence interval related to the population parameter that is mu 1. So, it is a confidence interval for mu 1 and the range is x 1 bar minus this quantity to x 1 bar plus this quantity, another issue what you want to do.

(Refer Slide Time: 09:29)

Random variable $(\bar{x}_1 - \bar{x}_2) \Rightarrow \mu_1 - \mu_2$

$$\leq \mu_1 - \mu_2 \leq$$

$$E(\bar{x}_1 - \bar{x}_2) = \mu_1 - \mu_2$$

$$V(\bar{x}_1 - \bar{x}_2) = V(\bar{x}_1) + V(\bar{x}_2)$$

$$= \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} = \frac{\sigma^2}{n} + \frac{\sigma^2}{n} = \frac{2\sigma^2}{n}$$

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

You may be interested to know that what is the interval estimation for \bar{x}_1 minus \bar{x}_m population minus m population that means what I mean to say, you will be interested to know μ_1 minus μ_m . μ_1 minus μ_m less than equal to this less than this interval, you want to find out and this one \bar{x}_1 minus \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_1 and μ_m , we will be using the statistics \bar{x}_1 minus \bar{x}_m . So, what you require to know, you require to the expected value of \bar{x}_1 minus \bar{x}_m this will be nothing but μ_1 minus μ_m .

You also require to know the variance component of \bar{x}_1 minus \bar{x}_m , this will be variance of \bar{x}_1 plus variance of \bar{x}_m and all of us know that this will be your σ^2/n_1 plus σ^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 σ^2/n_1 plus σ^2/n_m that is $2\sigma^2/n$ which is nothing but $2\text{MSE}/n$.

(Refer Slide Time: 11:34)

The image shows a handwritten derivation on a blue background. At the top right, there is a small logo for '© CET IIT KGP'. The main content consists of two lines of equations and a label:

$$\left(\bar{x}_1 - \bar{x}_m\right) - t_{N-L}^{(\alpha/2)} \sqrt{\frac{2\text{MSE}}{n}} \leq \mu_1 - \mu_m \leq \left(\bar{x}_1 - \bar{x}_m\right) + t_{N-L}^{(\alpha/2)} \sqrt{\frac{2\text{MSE}}{n}}$$

$$\left(\bar{x}_1 - \bar{x}_m\right) - t_{N-L}^{(\alpha/2)} \sqrt{\text{MSE}\left(\frac{1}{n_1} + \frac{1}{n_m}\right)} \leq \mu_1 - \mu_m \leq \left(\bar{x}_1 - \bar{x}_m\right) + t_{N-L}^{(\alpha/2)} \sqrt{\text{MSE}\left(\frac{1}{n_1} + \frac{1}{n_m}\right)}$$

Below the second equation, it is labeled "100(1-α) % CI" and "Bonferroni Approach". At the bottom left, there is a small NPTEL logo.

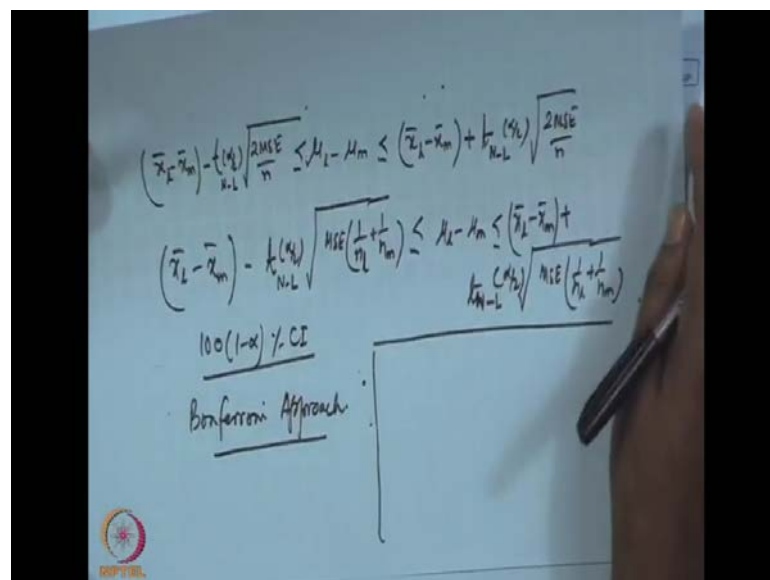
Now, what is my μ_1 minus μ_m , the interval then you must write down that this one, this is the first part then minus, we again consider the same t test. So, we will consider n minus $1 - \alpha$ by 2 only thing change is instead of MSE/n , it will be $2\text{MSE}/n$ see the 2 is coming earlier for one mean case, you have to use MSE $2\text{MSE}/n$. This side will be \bar{x}_1 minus \bar{x}_m minus, sorry plus $t_{n-1, \alpha/2}$ square root of 2

MSE by n. What will happen if your sample sizes are different? This quantity will not be 2 MSE by n exactly.

So, for unequal sample size case $\bar{x}_1 - \bar{x}_m \pm t_{n-1, \alpha/2} \sqrt{MSE \left(\frac{1}{n_1} + \frac{1}{n_m} \right)}$. You write second one n_1 with this one and m population less than equal to $\mu_1 - \mu_m$ less than equal to $\bar{x}_1 - \bar{x}_m \pm t_{n-1, \alpha/2} \sqrt{MSE \left(\frac{1}{n_1} + \frac{1}{n_m} \right)}$. Now, see we by seeing this, what we are saying this 100 into 1 minus alpha percent CI, correct. Is it true that it is 100 into 1 minus alpha percent CI? If I compare all the pairs, it will be reduced.

So, it is basically we are talking about 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. So, let us now again I will repeat one thing.

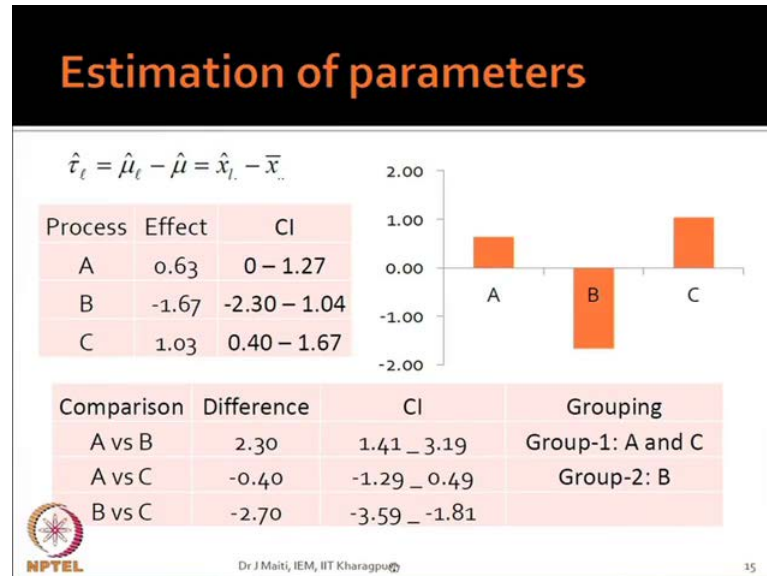
(Refer Slide Time: 14:50)



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. So, we will see here with an example.

(Refer Slide Time: 15:40)



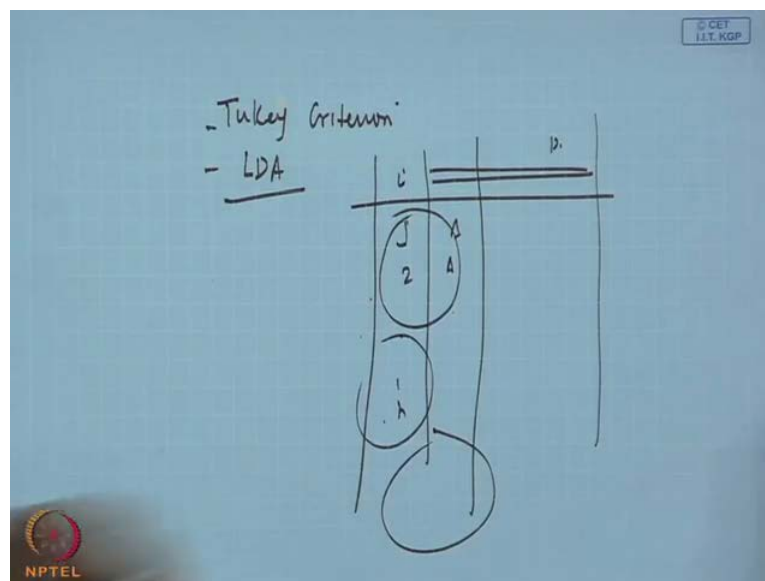
Now, see the same example we say that the process A B C the effect A's effect is 0.6 then B's effect is minus 1.67 and C's effect is 1.03. If you sum total, this what is happening how much it is all effect sum will be 0, keep in mind this one here it is basically rounding erode is there. So, 0.63 minus 1.67 1.03 that is 0.01 or something is coming, but it is rounding here then using the formulation, we have found out the confidence interval that is 100 into 1 minus alpha percent CI. Here, alpha 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 minus 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, 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 minus 0.40 and B to C minus 2.70 what are these, these are the mean differences. 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 confide 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 minus 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. 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. So, if you use Bonferroni approach, it will be a little tighter, but there is another many other techniques available.

(Refer Slide Time: 19:24)

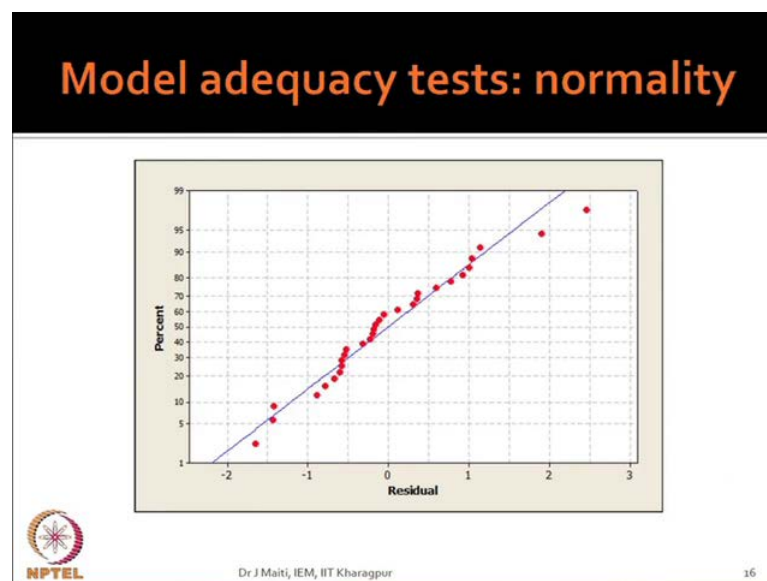


Like Tukey criteria, I think you know Fischer LDA linear discriminant analysis, which one this concept, which one you are talking about that grouping, that one and A and C. All those things clustering see ultimately you are some groups; you are getting essentially clustering means from the data you want 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 SSS values in term of samples from the processes. Essentially, you are not grouping the items in that sense 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.

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 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.

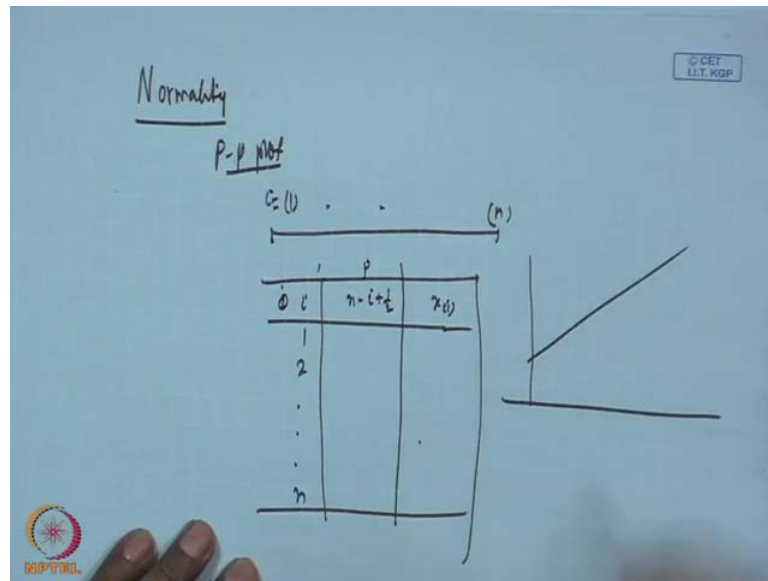
So, in clustering time you knew either has radical clustering or k means clustering C mean clustering that are there superior method. We should not go in that this manner, but again definitely I can say you have to inform of here two groups and that is why the third in this line that can. It be clustered that is the usage grouping is basically here we are saying from mean point of view they are similar.

(Refer Slide Time: 22:33)



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.

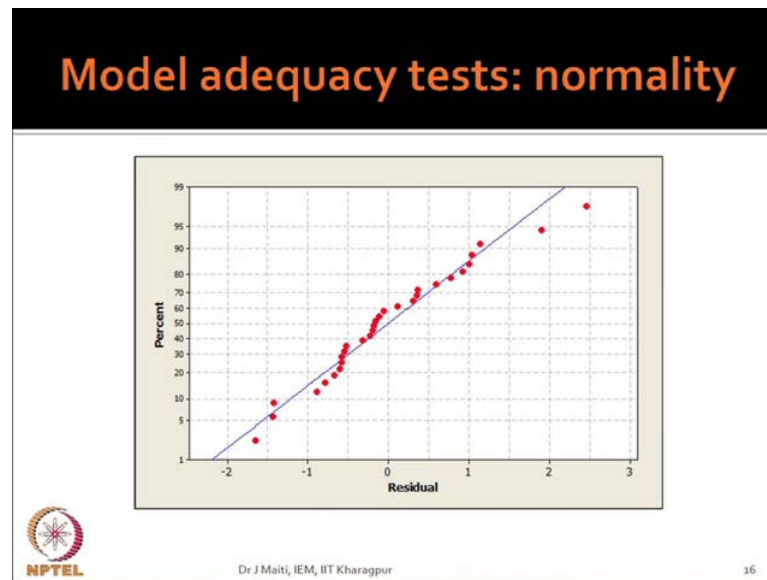
(Refer Slide Time: 22:58)



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 there, here what will happen in 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 have $n - i + 1$ plus half $n - i + 1$ plus half i equal to 1 2 to n.

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. You must get a straight line, this is probability plot here what is happening in this figure, and you see histogram what will happen?

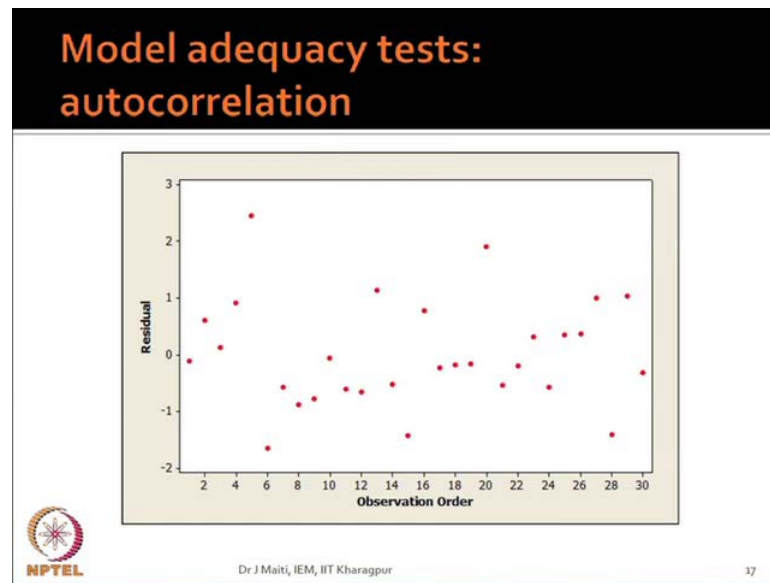
(Refer Slide Time: 22:58)



Then, you will not get that the from the histogram, you get an feel that that may be normally distributed, but you cannot say 100 percent whether it is normal or not. See 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 Kolmo graph seynough 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 cummulat depart percentage.

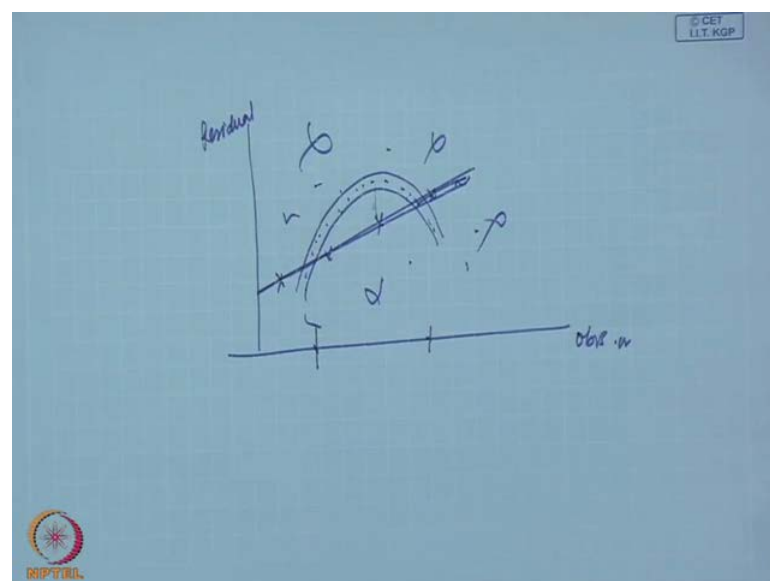
So, if really your n is 10 then the first one will be that is in the fifth percentile case rest will be 95 will be rest that thing. So, in that sense the far few fifth, tenth then 15 if number of observation is 10. So, in that way you are plotting here, so you will get a straight line if you do not get 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 last 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 that quantitatively. I think you can use z cartosis test is there then your K S test is 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, autocorrelation what is say observation should not be serially correlated. For example, first observation is related with the tenth observation right observation, that should not be correlated it may happen that over time. If 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 zero for example.

(Refer Slide Time: 27:44)

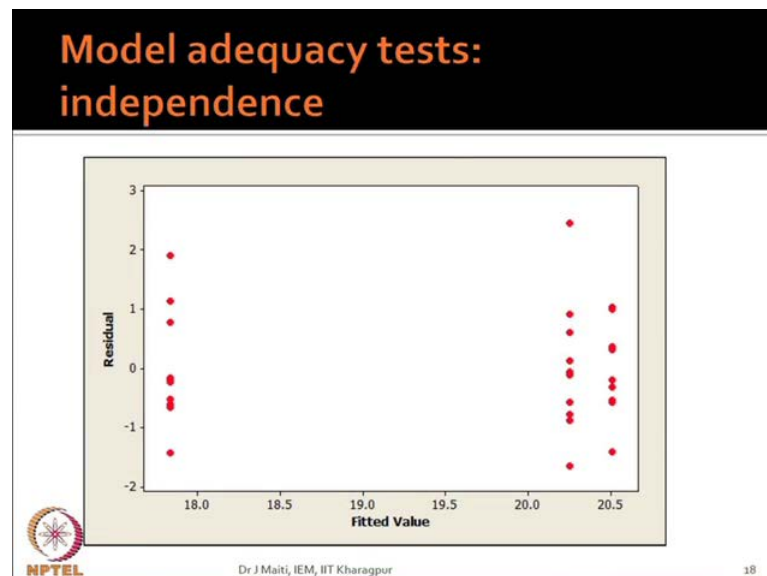


You are fitting a linear model like this my data is like this, so ultimately you will capture this much, but this nonlinear depart where it will go to the aero. So, in this aero dome that will be seen and another issue when I talk about that autocorrelation part when you talk about if there is correlation between this observations. This observation then 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.

So, if I see the aero observation order which is these the residuals there should not be any systematic pattern. In this case is there any pattern is you able to see that here is a linear or nonlinear or some increasing or decreasing trend. I do not think, it is a 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, 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 there this should not happen. It should be something like this nobody knows where it will be, so this says that our data is definitely independent then I think this is autocorrelation.

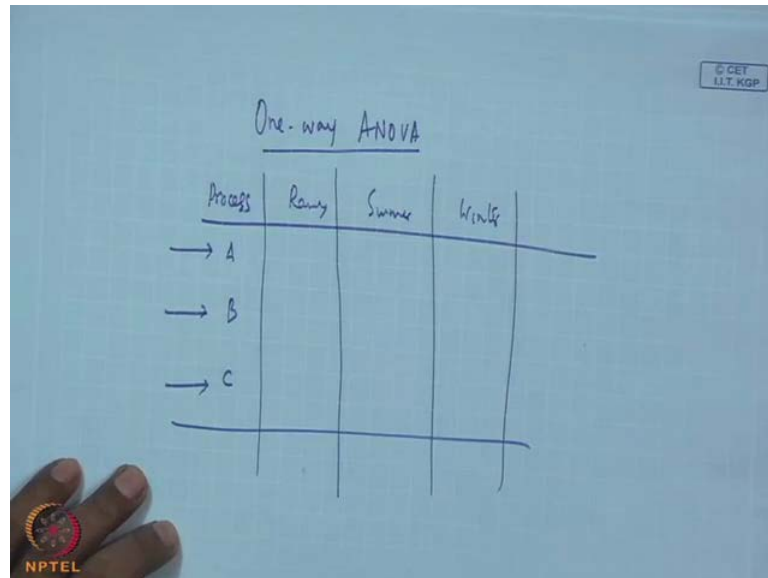
(Refer Slide Time: 29:53)



Then we go for independent test in independent test you are making the residual versus fitted value. Actually, what happened in this case same values for a particular fitted value, this is a fitted, this is a fitted value. 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: 30:56)



The image shows a hand-drawn table on a whiteboard titled "One-way ANOVA". The table has four columns: "Process", "Rainy", "Summer", and "Winter". The rows are labeled "A", "B", and "C" on the left side, with arrows pointing to each row. The table is currently empty, representing the structure for data collection or analysis. In the bottom left corner, there is a small circular logo with "NPTEL" written below it. In the top right corner, there is a small rectangular stamp that reads "© CEY I.I.T. KGP".

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 one population two and population three now what is the guarantee that 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: 32:05)

Process	Op1	Op2	Opk
A			
B			
C			

This is one second thing is, maybe suppose there are 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 mean with 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. Now, see you do not have control on the seasons, so even though you know, but you k you maybe some action plan you may be taking, but one over that seasons you have no control.

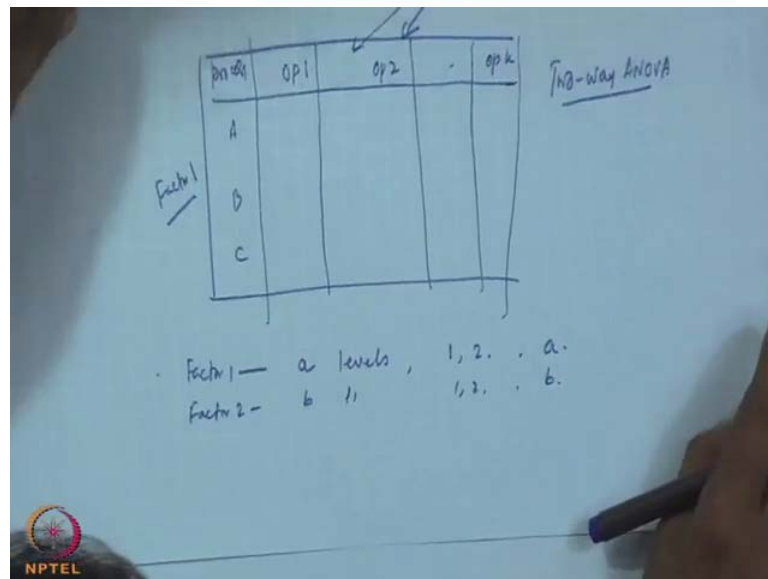
So, in that sense what will happen sometimes, suppose we can say that season is a noise variable, but you want to estimate the process of it keeping in view that the seasonal effects to be blocked what I am trying to say. So, that means you will go for any season production and see the quality summer 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. It may so happen that your control over the operators. So, depending on the situation you may change the operator. You give the operator to different machines where he maybe 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 effected too.

Even here, we can say noise variable this is factor 2 and process is factor 1, only difference between the first and second is blocking in this. 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 here you want to estimate both seasonal effect sorry operator's effect plus process effect. This type of design is two ways ANOVA; this is two way ANOVA, what will happen?

(Refer Slide Time: 35:35)



Here, why two way because factor 1 having suppose a levels that means factor 1 is a being factor 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: 36:01)

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

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. Usually, we denoted factor a or factor 1, factor b, a b is coming like this, now what will happen you will collect data. 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 \bar{x} . If I say this is \bar{x}_1 , I am giving one dot here what is this \bar{x}_1 , you can ignore factor 2, if you ignore factor 2 then the sum total of observations here this \bar{x}_1 , this \bar{x}_1 , this \bar{x}_1 , this \bar{x}_1 , this observation is related to 1. So, you will be getting one average, you are ignoring that too, similarly here you are getting two 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 $\bar{x}_{.j}$, $\bar{x}_{.j}$ stands for all the rows here that are levels for factor 1 $\bar{x}_{.j}$, $\bar{x}_{.j}$ stands for levels in factor 2. So, then this will be $\bar{x}_{1.}$, $\bar{x}_{2.}$, so like this I think this one be $\bar{x}_{b.}$, then here will one grand average, 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 dot this minus that, similarly here also we can find out y , I think we will not use this one dot, we will not use here τ_1 . Similarly, here we will use τ_m , then this one will be your $\bar{x}_{1.}$ minus $\bar{x}_{.2}$ minus $\bar{x}_{.}$ like this $\bar{x}_{.b}$ minus $\bar{x}_{.}$. This is known as two ways ANOVA, two way one factor 1 l 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

N = abn

A source of variation one is factor 1 factor 2 is also urge interaction one and two 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, factor 2, sorry if I write like this SS 1 and SS 2 no problem, then factor 2 then SS 1 2, 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. So, here what will be the degrees of freedom, how many levels are there for factor 1 a level, so a minus 1 how many levels are there for factor 2, b minus 1 what will be the interaction, a minus 1 b

minus 1, what is the total number of observation a b n minus 1. 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 a b cells and in each cells there are n observations a b n, what is our capital. Now, if you collect different samples for different shells and as well as if you collect different samples for the factor 1 and factor 2 differently, ultimately this computation will be little different. You can still write n, you have to compute n then what will be the SSC space a b n minus 1, a minus 1 minus b minus 1 minus a minus 1 into b minus 1, because total will be this.

(Refer Slide Time: 44:01)

The image shows a hand writing the following derivation on a whiteboard:

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

The whiteboard also features an NPTEL logo in the bottom left corner and a small box in the top right corner containing the text "© CET I.I.T. KGP".

So, what will be this value a b n minus 1, a minus 1 minus b minus 1 minus a minus 1 b minus 1, so this a b n minus 1 minus a plus 1 minus b plus 1, so minus this is a b minus a minus b plus 1. So, I am writing a b n minus 1 minus a plus 1 minus b plus 1 minus a b plus a plus b minus 1, so plus a plus minus a plus b minus b minus 1 plus 1, then it is basically this is what minus 1, this is plus 1, this also cancelled out, so that mean a b n minus 1.

(Refer Slide Time: 45:08)

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 (12)	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

Then, what you require to know you require to compute MS mean square that variability this will be MS 1 will be SS 1 by a minus 1 MS 2 will be SS 2 by b minus 1 then MS 1 2 that is interaction which is SS 1 2 by a minus 1 into b minus 1. Then your MSE will be SSE by AB into n minus 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 if 1 F 1 if I say for the factor 1, this one will be MS 1 by MSE for factor 2 it will be MS 2 by MSE for 1 2 it is MS 1 2 by MSE. These are the things you want to test, 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 minus 1 and degree of freedom is ab into n minus 1. If this alpha this is the case, so if computed F is greater than the tabulated F then what will be your decision reject H₀. So, similarly if this 1 F b minus 1 a b n minus 1 do it if this 1 F a minus 1 into b minus 1, I think into a minus 1 into b minus 1 and a b n minus 1. Then do this is what is known as two way ANOVA.

(Refer Slide Time: 48:31)

3-Way ANOVA

Factor A — a levels
Factor B — b levels
Factor C — c levels

Sample size = n,
 $N = abc n$

Sources of Variation	
A	
B	
C	
AB	
AC	
BC	
ABC	

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 3 way ANOVA that mean if I say factor. Now, writing in terms of A B C 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 a b c n. 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? 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.

(Refer Slide Time: 49:50)

Factor C -

Source / Variability	SS	DOF	F
→ A	SSA		F_A
→ B	SSB		F_B
→ C	SSC		
→ AB	SSAB		
→ AC	SSAC		
→ BC	SSBC		
→ ABC	SSABC		
→ Error	SSE		
	SST		

Handwritten notes on the table:

- Main effects: A, B, C
- 2-way interaction effects: AB, AC, BC
- 3-way interaction effects: ABC

See, first is A, second is B, third is C, AB may interact BC may interact AC 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 minus 1, b minus 1, c minus 1, inter interaction of it a minus 1 into b minus 1 for AB, for AC A minus 1 into c minus 1. Like this, ABC case A minus 1, B minus C minus that multiplication and your toe for your total case that will be ABC n minus 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 we have to compute it F a F b like this.

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 in 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 you are in a position to explain ANOVA for any number of factors just you just you compute little bit.

For three digit, 2 to the power 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 n sa 1 is coming, then two at a time, n c 2 is coming. So, that is three factors two 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, ask me some question. So, this is basically the totality from ANOVA point of view, I think ANOVA is very useful, 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 and for ANOVA, I suggest you dc Montgomery design analysis of experiment.

(Refer Slide Time: 54:36)



The slide features a black header with the word "References" in orange. Below the header, a white box contains a list of two references. At the bottom left of the white box is the NPTEL logo, and at the bottom center is the text "Dr. J. Maiti, IEM, IIT Kharagpur". At the bottom right of the white box is the number "19".

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.

 NPTEL

Dr. J. Maiti, IEM, IIT Kharagpur

19

This is the best book I have read in terms of your ANOVA, large number of real life examples is 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.

Thank you very much.