Business Analytics for Management Decision

Prof. Rudra P Pradhan
Vinod Gupta School of Management
Indian Institute of Technology, Kharagpur

Lecture – 24
Inferential Analytics Part – 2 (Contd.)

Hello everybody. This is Rudra Pradhan here and welcomes you all to BMD lecture series and today we will continue the inferential analytics in that to part 2 and we are on the process of discussing multiple sample case. In the last couple of lectures we have discussed various cases by using t statistics, z statistics and F statistics. Again we will discuss the same problems by using F statistics and that too in multiple variate, means multiple sample case. So, typically it is a multivariate case.

So, when we have actually one sample then the problem is very simple ones then little bit complex you may have a problem with it 2 samples and then well again the complexity we will start with when we have a more sample simultaneously. So that means, the moment will be moved from univariate to multivariate then the complexity will start increasing.

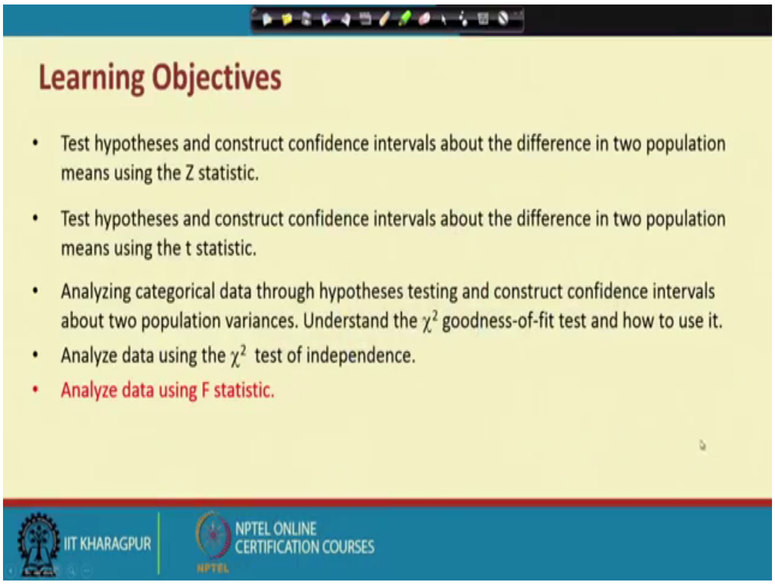
So, as a result, the test structure will be slightly simple to complex. So, what will it do here is, we like to take a particular you know business problem and we will see when there are multiple sample case what are the process through which you have to what are the process through which you have to check the kind of know difference. So, now, if it is actually one sample case, we are checking that you know whether sample mean is you know having differ difference with the population mean. And when there are 2 sample case then the mean of the 2 samples the difference of 2 mean samples is a converging towards you know population mean difference and something like that.

So, now when we have a multiple sample case, typically you know let us say there are you know 4 variables and for; that means, equally 4 sample case and then we have a you know different kind of observations. And then in the multiple sample case one of the typical inference which you would like to have or you like to observe is corresponding to this problem and sample of observations. So, we like to check whether you know there is a significant difference within the samples and between the samples.

So, again we will go for some kind of variance check and that the consistency measure and here the F statistic is to check whether the difference between the variance between the samples and within the samples are significantly different as different to you know population kind of variance. So, here the idea is to check the mean difference between you know within the group and between the group.

So, what are the process through which you have to investigate let us see you know the case of the structure called you know ANOVA.

(Refer Slide Time: 03:20)

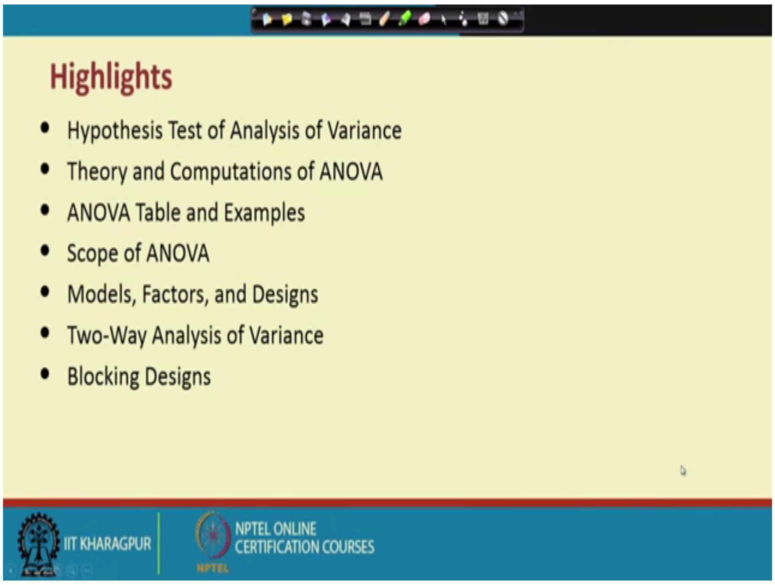


Learning Objectives

- Test hypotheses and construct confidence intervals about the difference in two population means using the Z statistic.
- Test hypotheses and construct confidence intervals about the difference in two population means using the t statistic.
- Analyzing categorical data through hypotheses testing and construct confidence intervals about two population variances. Understand the χ^2 goodness-of-fit test and how to use it.
- Analyze data using the χ^2 test of independence.
- Analyze data using F statistic.

So, these are the learning objectives which you are you know having and the idea I means the highlights of this lecture is to typically go for analysis of variance.

(Refer Slide Time: 03:25)



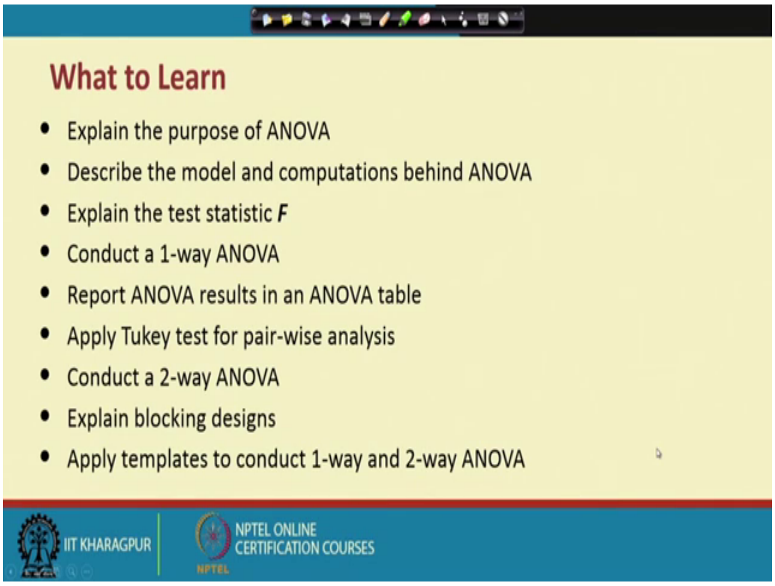
Highlights

- Hypothesis Test of Analysis of Variance
- Theory and Computations of ANOVA
- ANOVA Table and Examples
- Scope of ANOVA
- Models, Factors, and Designs
- Two-Way Analysis of Variance
- Blocking Designs

Then that to we would like to know the theoretical aspects or computation of ANOVA that is the analysis of variance and then ANOVA tables and examples. Of course, we have already discussed the concept of ANOVA tables with the examples.

But here the process is a little bit you know complex and that too with multiple sample case then scopes a scope of ANOVA and the extent was extended version of ANOVA that to models factors and designs then 2-way analysis of variance and then finally, block designs.

(Refer Slide Time: 04:11)



What to Learn

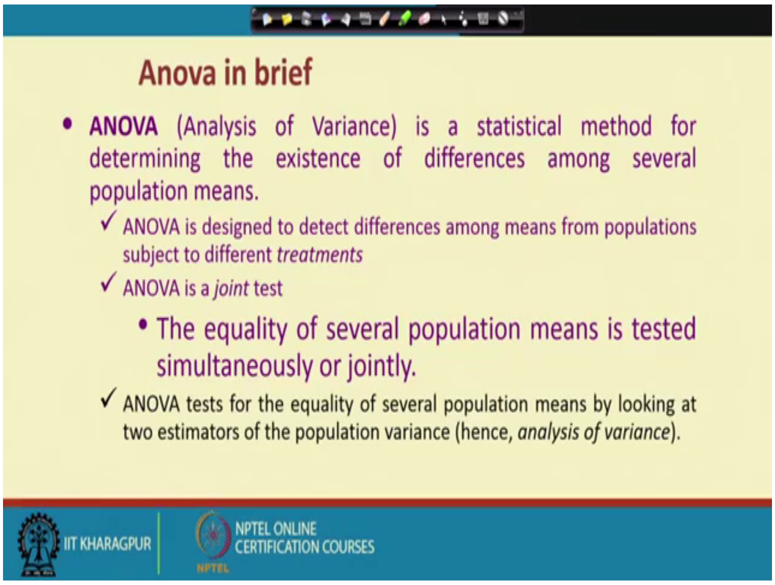
- Explain the purpose of ANOVA
- Describe the model and computations behind ANOVA
- Explain the test statistic ***F***
- Conduct a 1-way ANOVA
- Report ANOVA results in an ANOVA table
- Apply Tukey test for pair-wise analysis
- Conduct a 2-way ANOVA
- Explain blocking designs
- Apply templates to conduct 1-way and 2-way ANOVA

So, the learning idea behind this particular structure is we like to know what is the explanations behind the purpose of proper of ANOVA and describes the model and computations behind ANOVA explained it you know test statistic of a ANOVA and we will connect with a 1-way ANOVA, 2-way ANOVA, then 3 way ANOVA and that is the concept called as ANOVA.

So, that is why what I have mentioned earlier. So, it is the degree of complexity. So, typically if you will go by you know F statistic only. So, then the complexity we will start with you know 1-way ANOVA, 2-way ANOVA, 3 way ANOVA then the concept called as a MANOVA that is called as a multivariate analysis of variance. So, now, in the work 1-way ANOVA case with respect to a particular attribute we will check the difference and in the case of 2-way ANOVA we may have a 2 different layers through which you check the difference and in the case of MANOVA you have a multiple layers through which you have to check the difference.

So, let us see first you know the structure with respect to 1-way ANOVA.

(Refer Slide Time: 05:15)



Anova in brief

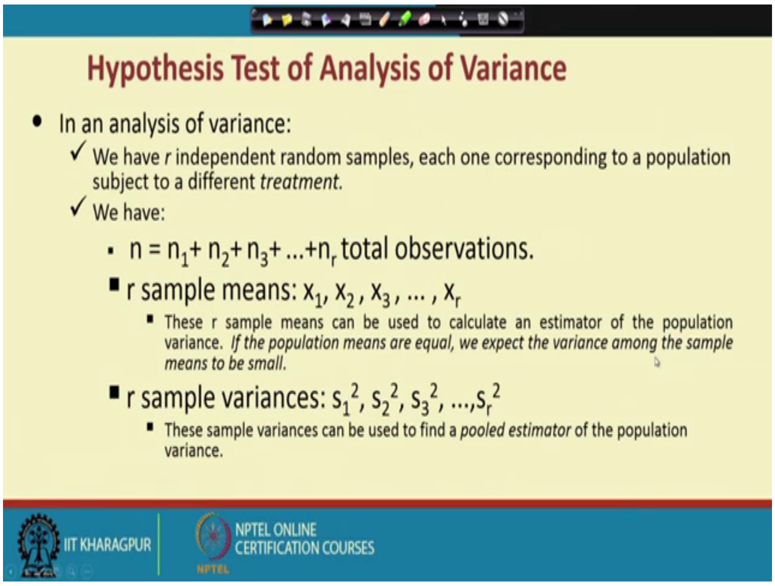
- **ANOVA** (Analysis of Variance) is a statistical method for determining the existence of differences among several population means.
  - ✓ ANOVA is designed to detect differences among means from populations subject to different *treatments*
  - ✓ ANOVA is a *joint* test
    - The equality of several population means is tested simultaneously or jointly.
  - ✓ ANOVA tests for the equality of several population means by looking at two estimators of the population variance (hence, *analysis of variance*).

Before you start the process of 1-way ANOVA let me highlight what is exactly the ANOVA concept and of course, we have already highlighted little bit about this concept in the last lecture. So, ANOVA that is an analysis of variance is a statistical method for determining the existence of differences among several population means. So, that means here actually we have a multiple samples and each sample is drawing from different populations.

So, that is why we have a different population means and we have to, we have also different samples and will you draw different samples and then we will check if in a different sample variance with you know population variance that is you or the idea behind this particular you know structure what we call as you know ANOVA. And ANOVA is designed to detect the difference among means from populations subject to different treatments and it is a joint test. The quality of several population means is tested simultaneously or we can say that you know jointly.

ANOVA test for the equality of several population means by looking at the 2 estimators of the population variance and hence it is the kind of analysis of variance. Typically it is a variance statistics and a any particular in a particular situation or any particular you know k test or t test the calculated value may be a negatively skewed may be positively skewed but in this case it will be it is always actually positively loaded positively skewed.

(Refer Slide Time: 06:33)



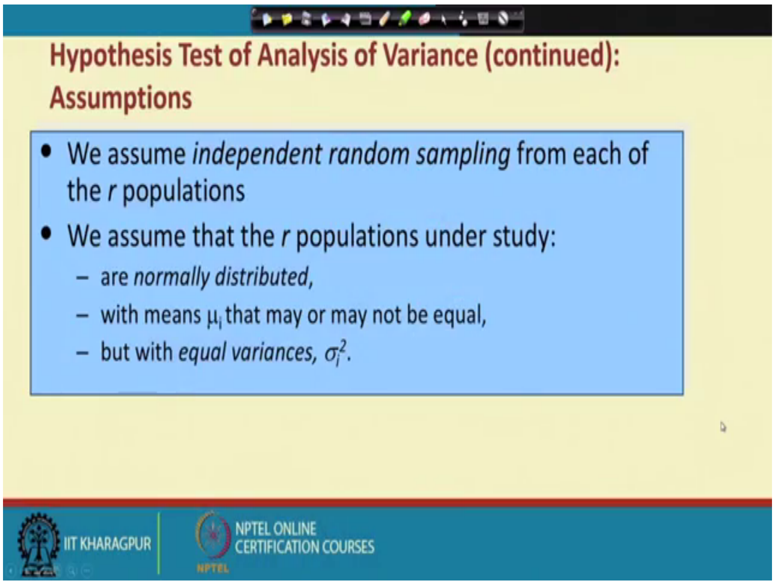
Hypothesis Test of Analysis of Variance

- In an analysis of variance:
  - ✓ We have *r* independent random samples, each one corresponding to a population subject to a different *treatment.*
  - ✓ We have:
    - $n = n_1 + n_2 + n_3 + \ldots + n_r$ total observations.
    - r sample means: $x_1, x_2, x_3, \ldots, x_r$
      - These r sample means can be used to calculate an estimator of the population variance. *If the population means are equal, we expect the variance among the sample means to be small.*
    - r sample variances: $s_1^2, s_2^2, s_3^2, \ldots, s_r^2$
      - These sample variances can be used to find a *pooled estimator* of the population variance.

In an analysis of variance we have r independent random samples and each one corresponding to a population subject to a different treatment and ultimately the total sample size will be n 1 plus n 2 plus n 3 up to you know n rth observations. So, that is what the total populations.

Then the samples which we are drawing a you know r samples which you are drawing from the population n that is x 1 x 2 x 3 and then sample variance corresponding to x 1 x 2 x 3 are s 1 square s 2 s 2 square s 3 square and s r square. So, these are the background about this ANOVA, that means actually we have a multiple populations case and we are doing you know individual samples from these populations. And then we are checking the mean of these samples or variants of these samples are statistically different from each other whether there is a difference or there is no difference depending upon the kind of inference we will take some kind of management decisions. So, that is we are here to test the particular you know structure.

(Refer Slide Time: 07:51)

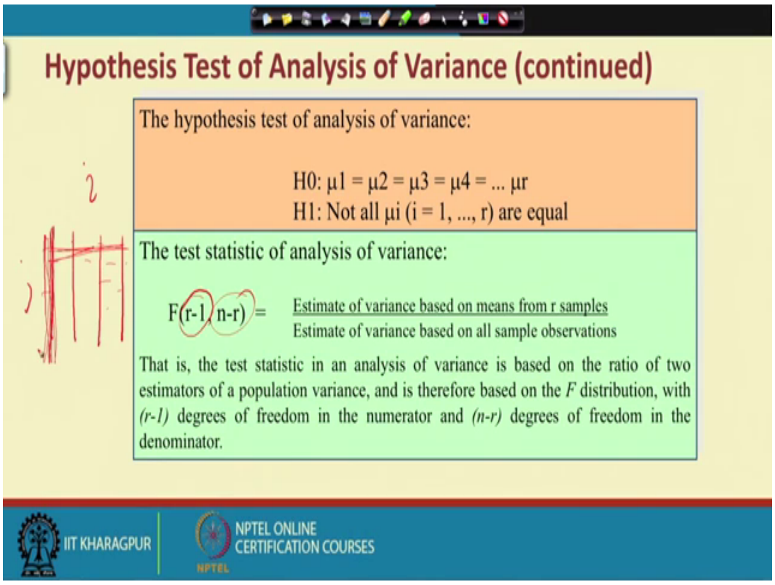


Hypothesis Test of Analysis of Variance (continued): Assumptions

- We assume *independent random sampling* from each of the *r* populations
- We assume that the *r* populations under study:
  - are *normally distributed*,
  - with means μ_i that may or may not be equal,
  - but with *equal variances*, σ_i^2.

So, hypothesis t test for ANOVA you know having some kind of assumptions. So, we assume that you know independent random sampling from each of the rth populations and we again assume that the r population under the umbrella we know they are you know normally distributed and with means mu I that may or may not be equal and then with equal variance. So, that means typically we have r populations and we are drawing samples and that you know then mean of the samples, we are assuming that you know mean of the samples are you know equal, but we are checking the kind of difference with respect to equal variance that is actually sigma square i, right.

(Refer Slide Time: 08:40)



Hypothesis Test of Analysis of Variance (continued)

The hypothesis test of analysis of variance:

H0: $\mu 1 = \mu 2 = \mu 3 = \mu 4 = \ldots \mu r$

H1: Not all μi ($i = 1, \ldots, r$) are equal

The test statistic of analysis of variance:

$$F(r\text{-}1, n\text{-}r) = \frac{\text{Estimate of variance based on means from r samples}}{\text{Estimate of variance based on all sample observations}}$$

That is, the test statistic in an analysis of variance is based on the ratio of two estimators of a population variance, and is therefore based on the *F* distribution, with *(r-1)* degrees of freedom in the numerator and *(n-r)* degrees of freedom in the denominator.

So, with this particular conditions the hypothesis testings for this ANOVA is like this. So, we are testing that you know whether the sample means are you know different to each other. Since we have a multiple samples, here the idea is actually a corresponding to multiple samples and they are drawn from the different populations. So, their population means are you know equal that is the hypothesis we have to build and corresponding to this hypothesis null hypothesis the alternative hypothesis that they are not equal.

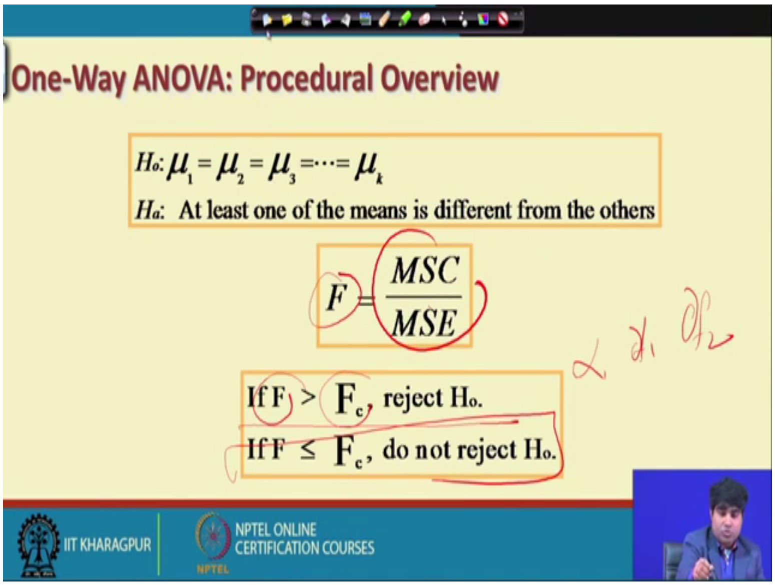
So; that means, mu i equal to you know mu i equal to you know you know mu 1, mu 2, mu 3 like this or mu i not equal to you know for all I starting from 1 2 you know r th row or rth number of series. So, that means typically in simple language. So, all the means are you know equal populous means are equals and population means are not equal this is the simple way you can elaborate and then you draw the samples and check whether there is a significant difference between the a 2.

So, your here F statistic follows the ratio between estimate of variance based on means of means from the r th samples and divided by estimate of variance based on all sample observations. So, that is how the again between you know on the kind of upper part and the lower parts. So, we are checking the kind of difference and that is the significance difference between these 2 groups.

And then since you know we have already discussed the ANOVA concept in the last lecture for F critical. So, we need 2 degree of freedoms and in this particular case and in this particular case. So, here you know first degree of freedom will be r minus 1 and n minus r. So, you know question actually the typical ANOVA structure will be like this. So, this is the series and here let us say i series and j series. So, we like to check whether you know. So, these are the points these are the points. So, we like to check whether you know there is a difference between within the group and between the group.

So, now, these 2 difference should not be statistically significant and whether they should not be significant or not then depending upon you know samples or the kind of problem so we have to check. So, if there is a significant difference between these 2, your you know the kind of decision the kind of strategy will be different, but we need to know the kind of inference and on the basis of this inference we have to take a decisions what to, what should be the next step behind the particular you know on the basis of this particular you know problems.

(Refer Slide Time: 11:25)



One-Way ANOVA: Procedural Overview

$H_0: \mu_1 = \mu_2 = \mu_3 = \cdots = \mu_k$

H_a: At least one of the means is different from the others

$$F = \frac{MSC}{MSE}$$

α, df_1, df_2

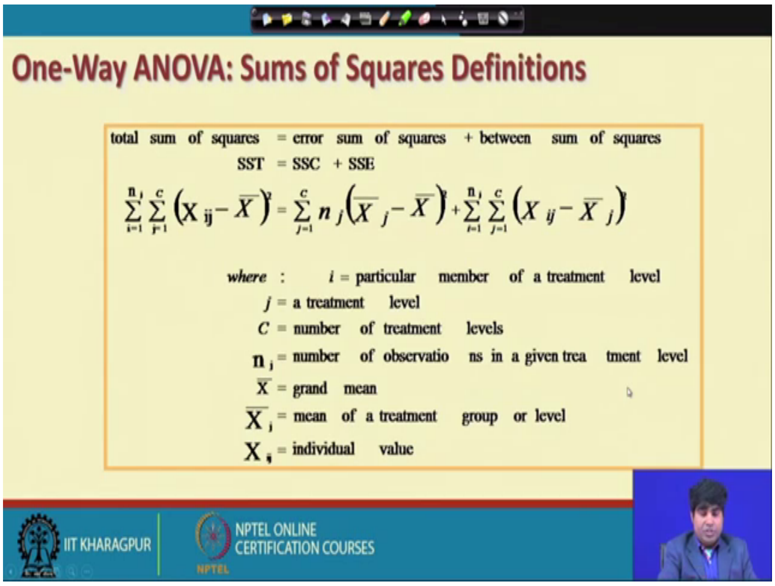
If F > F_c, reject H_0.

If F ≤ F_c, do not reject H_0.

So, now typically the 1-way ANOVA structure is like this which you have already highlighted. So, all the population means are equal and at least one of them are you know different from others, right. Maybe actually the suppose there are you know for populations mu 1, mu 2, mu 3, mu 4 at least between 2 pairs are not actually equal that is the you know conditions for alternative hypothesis. And the testing procedure follows the F statistics and it is the ratio between these 2 mean squared you know statistics and F you know the kind of outcome or the kind of inference typically depends upon the comparison between F critical and F calculated.

So, typically we will we will calculate F like this and then I will discuss details about this particular you know structuring in the next slide. So, in the mean time, this is F calculated and the decision rule will be when F calculators will be overtaking the F criticals depending upon the alpha then df 1 and df 2 and then if not then we cannot reject the hypothesis if you you know the F calculator will be overtake the F critical then we have to reject the null hypothesis. That means, in this case, your null hypothesis mu 1, mu 2, mu 2 are not actually equals. So, that is what the inference or the kind of conclusion which you draw on the basis of this particular testings.

(Refer Slide Time: 12:58)



One-Way ANOVA: Sums of Squares Definitions

total sum of squares = error sum of squares + between sum of squares

SST = SSC + SSE

$$\sum_{i=1}^{n_j}\sum_{j=1}^{c}\left(X_{ij}-\overline{X}\right)^2=\sum_{j=1}^{c}n_j\left(\overline{X}_j-\overline{X}\right)^2+\sum_{i=1}^{n_j}\sum_{j=1}^{c}\left(X_{ij}-\overline{X}_j\right)^2$$

where : i = particular member of a treatment level

j = a treatment level

C = number of treatment levels

n_j = number of observatio ns in a given trea tment level

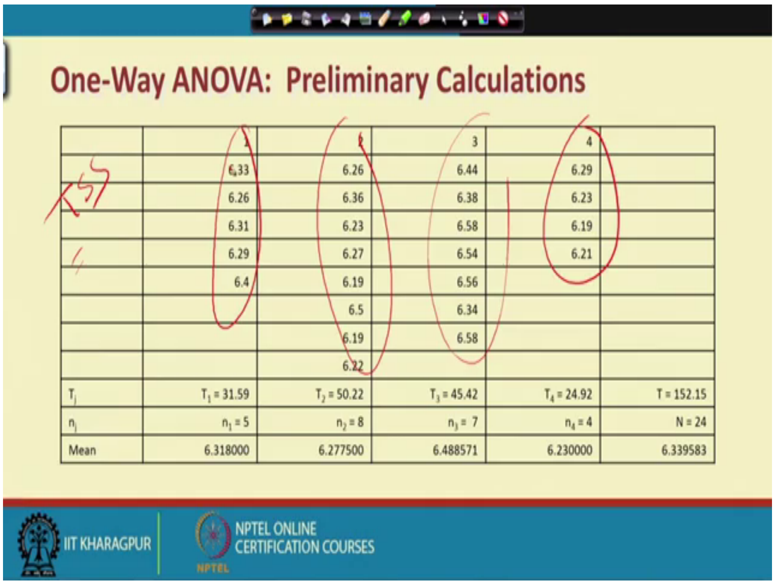
$\overline{X}$ = grand mean

$\overline{X}_j$ = mean of a treatment group or level

X_{ij} = individual value

Let us see how is this particular you know testing. So, here actually in the entire analysis of this particularly in ANOVA, so the typical structure will be. So, let me give you a sample problem first then I will come back again here. So, the typical examples will be like this and here you see here. So, this is actually one sample this is your second sample then third sample and 4 samples.

(Refer Slide Time: 13:16)



One-Way ANOVA: Preliminary Calculations

	1	2	3	4	
TSS	6.33	6.26	6.44	6.29	
	6.26	6.36	6.38	6.23	
	6.31	6.23	6.58	6.19	
	6.29	6.27	6.54	6.21	
	6.4	6.19	6.56		
		6.5	6.34		
		6.19	6.58		
		6.22			
T_j	T_1 = 31.59	T_2 = 50.22	T_3 = 45.42	T_4 = 24.92	T = 152.15
n_j	n_1 = 5	n_2 = 8	n_3 = 7	n_4 = 4	N = 24
Mean	6.318000	6.277500	6.488571	6.230000	6.339583

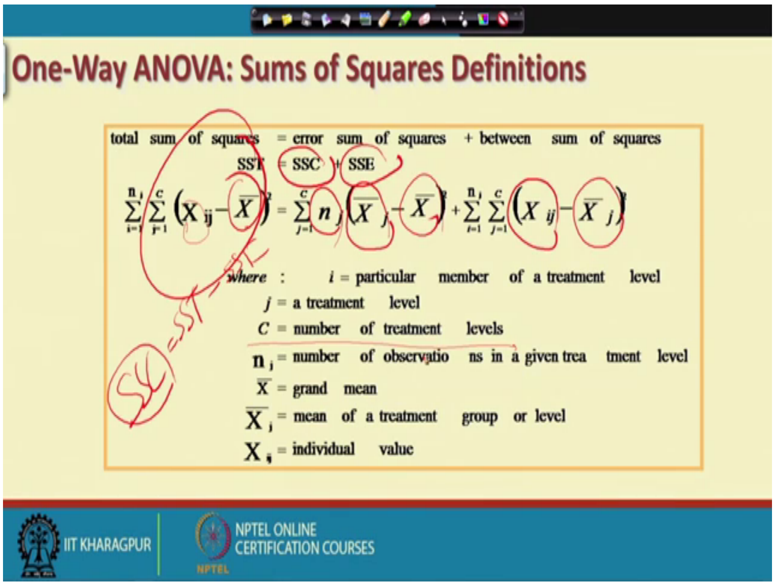
IIT KHARAGPUR NPTEL ONLINE CERTIFICATION COURSES

So, now, not necessarily that you know they have a uniform observation they may have a different observations and the idea behind this null hypothesis that you know it is drawn from the population 1, population 2, population 3, population 4 and then the a typical null hypothesis that you know mu 1 not equal to mu 2 not equal to mu 3 not equal to mu 4 and at least one pair should not be actually equal that is how the testing is all about.

So, we are in the process of you know we have a total samples by you know clubbing all these things. So, as a result, there are 3 typical variants we defined in this particular you know process t you know that is what the ANOVA all about. So, that is called as you know total sum of squares by using all the samples and we are checking the in that is actually its nothing, but nothing, but actually the difference between the square of the difference between the sample point individual sample point with the combined mean.

So, that is what the total variance that we called as you know total variance. We should be equal to the variance within the group and variance between the groups. So, that is the typical you know and the kind of F is nothing, but actually the mean square within the group the ratio between mean square within group and the mean square between the group and followed by the corresponding degree of freedom. So, let us first you know discuss the particular concept and then we will come back to this particular you know problems. So, accordingly, the typical structure will be like this. So, we can, this is how the particular structures what we have already discussed and there are you know all 3 variants variances.

(Refer Slide Time: 14:58)



One-Way ANOVA: Sums of Squares Definitions

total sum of squares = error sum of squares + between sum of squares

SST = SSC + SSE

$$\sum_{i=1}^{n_j}\sum_{j=1}^{c}\left(X_{ij}-\overline{X}\right)^2=\sum_{j=1}^{c}n_j\left(\overline{X}_j-\overline{X}\right)^2+\sum_{i=1}^{n_j}\sum_{j=1}^{c}\left(X_{ij}-\overline{X}_j\right)^2$$

$SY = SST - SE$

where : i = particular member of a treatment level

j = a treatment level

C = number of treatment levels

n_j = number of observatio ns in a given trea tment level

$\overline{X}$ = grand mean

$\overline{X}_j$ = mean of a treatment group or level

X_{ij} = individual value

So, this is actually sum square total that is total variance within the particular you know combined samples you know including all you know individuals and; that means, n 1, n 2, n 3. So, for instance corresponding to that previous problem, so we have a 4 samples.

So, the total samples will be n 1 plus n 2 plus n 3 plus n 4 and then also sum up all the items divided by the corresponding sample points will give you the mean that is called as you know combined mean and the variance you know the individual means become individual you know individual value of x corresponding to combined mean and then squaring if you add up for all the samples then the particular variance which is called as you know sum SST that is a total sum of square or sum square total and then it will be equal to exactly error sum of squares that and then between sum of squares. So, that means within the difference and between the difference.

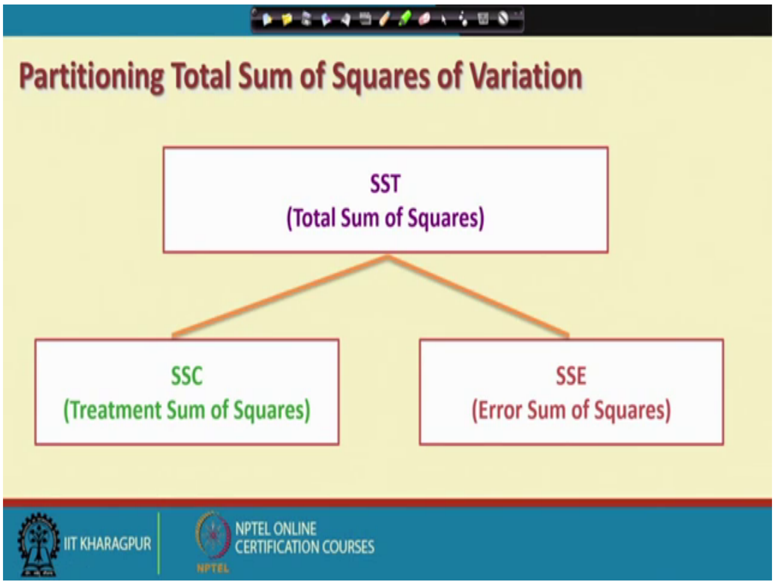
So, that is typically represented by SSC and SSE. So, I mean is mean S E and mean S C is nothing, but actually SSE divided by the corresponding degree of freedoms and the a corresponding degree of freedom and accordingly you will calculate the F statistics. And in the SSC its nothing, but actually this is the mean of a particular sample and this is the combined means and this is the sample adjustment. So, then in the case of between sum of square. So, it is the sample of a particular you know variables and then the mean of that particular you know variables.

So, in any particular point of time you know you need to calculate only 2 items and in fact, actually SST equal to SSC and applause you know SSE this is always true for any problem or any contest. So, as a result you know if you like to report manually then at a particular point of time at least 2 you can you know report then by default third one can be obtained. So, for instance let us just see, if you need SSC then you need to calculate SST first and then SSE. So, by default you can you know difference between the 2 you can get you know SS C or if you directly connect with the software, software by default will give you the values of all these you know you know items that is SST SSE and SSE. So, SSC plus SSC will give you the weight of SST.

So, accordingly, these are the following descriptions C, C is the number of treatment levels. So, that is actually column means it is typically the structure is the column wise variations and a row wise variations. So, this is the arrangement and sometimes say the problem can be analyzed with the balanced samples and sometimes the problem can be analyzed with you know unbalanced samples, not necessarily the sample size will be uniform in all the cases like the component which you have discussed earlier like correlation coefficient. In this case there is a difference the difference of sample size may not affect, but ultimately we need to know what is the significant difference between within the variance and between the variance.

So, now, corresponding to these, let us check what is the exactly concept.

(Refer Slide Time: 18:30)



Partitioning Total Sum of Squares of Variation

SST
(Total Sum of Squares)

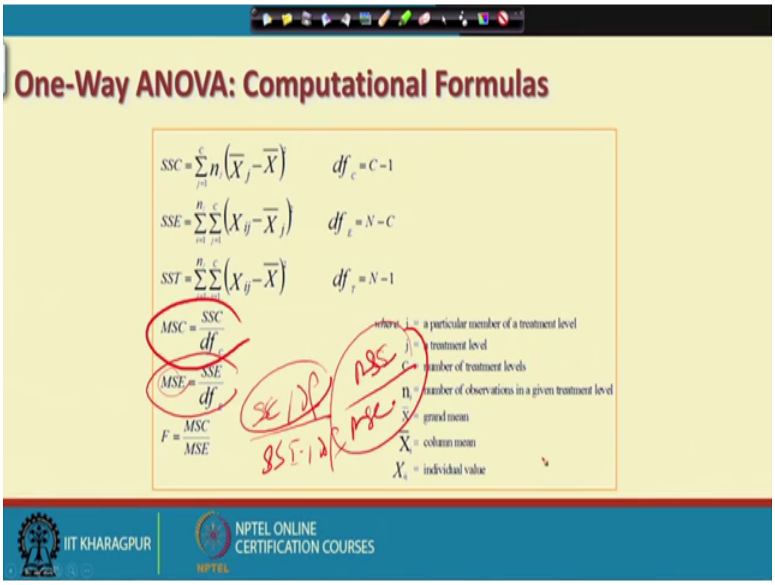
SSC
(Treatment Sum of Squares)

SSE
(Error Sum of Squares)

So, what we have already discussed. So, SST equal to SSC plus SSC that is you know nothing, but total sum of square equal to treatments sum of squares and error sum of squares. So, this is a different kind of concept, but it is the typical actual structure is a total variance is nothing, but actually within variance and between variance. So, this is this is the other way you can represent the particular. So, total variance sum of squares total means it is the total variance you know for the combined samples then you would like to check actually a within variance for individual samples then between variance for each case or each curve you know kind of row.

So, as a result you will get you know the concept called as you know sum square totals. So, corresponding to these particular structures, so the ANOVA table will be like this.

(Refer Slide Time: 19:25)



One-Way ANOVA: Computational Formulas

$SSC = \sum_{j=1}^{C} n_j (\bar{X}_j - \bar{X})^2 \qquad df_C = C - 1$

$SSE = \sum_{i=1}^{n_j} \sum_{j=1}^{C} (X_{ij} - \bar{X}_j)^2 \qquad df_E = N - C$

$SST = \sum_{i=1}^{n_j} \sum_{j=1}^{C} (X_{ij} - \bar{X})^2 \qquad df_T = N - 1$

$MSC = \frac{SSC}{df_C}$

$MSE = \frac{SSE}{df_E}$

$F = \frac{MSC}{MSE}$

where: i = a particular member of a treatment level

j = a treatment level

C = number of treatment levels

n_j = number of observations in a given treatment level

$\bar{X}$ = grand mean

$\bar{X}_j$ = column mean

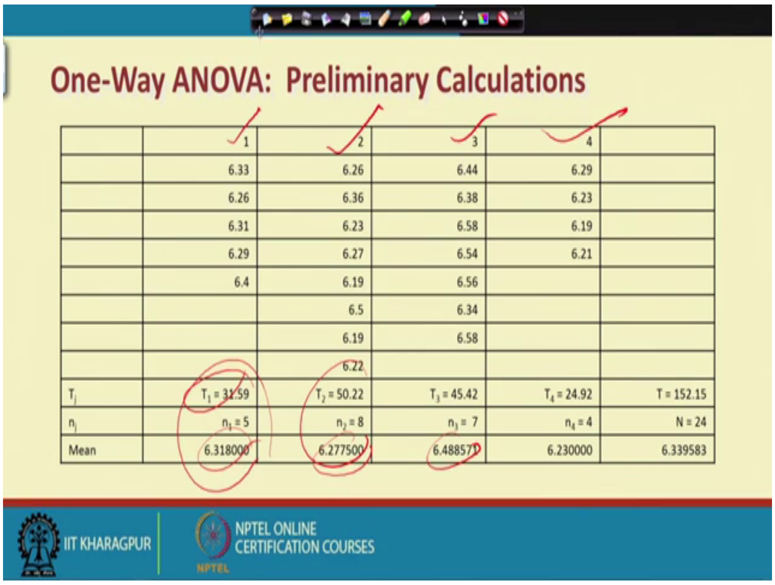
X_{ij} = individual value

So, first you calculate the sum SSC and then a then the kind of then the kind of SSE and then finally, SST. So, either you can individually calculate and again for simplicity you can calculate any 2 and the third one can be of obtained you know by default. And the adjustment factor will be here, so MSC is nothing, but SSE by kind of difference and. So, this is what actually. So, this is the SSE by degree of freedom and this is the SSC by degree of freedom this means sum squares and F is the F is nothing, but actually. So, it is the SSC a SSC by degree of freedoms and then and then SSE by degree of freedom.

So, this is nothing, but actually called as you know MSC and this is nothing, but called as you know MSC and this is what actually the a concept called as you know F calculated. And then you know with the help of samples, you can get the F value there on the basis of F value. So, we have actually a degree of freedom one and degree of freedom 2 that is actually sample adjustment row wise and column wise then finally, you have to check you know what is the F critical. And the final check will be or the decision making process in the or the decision will be the comparison between F calculated versus F critical at a particular probability levels that is depends upon the alpha which you have to fix or you have to or focus.

Accordingly, the typical structure will be like this let us take with you know discuss with your kind of an examples, so the same example which I have highlighted here.

(Refer Slide Time: 21:17)



One-Way ANOVA: Preliminary Calculations

	1	2	3	4	
	6.33	6.26	6.44	6.29	
	6.26	6.36	6.38	6.23	
	6.31	6.23	6.58	6.19	
	6.29	6.27	6.54	6.21	
	6.4	6.19	6.56		
		6.5	6.34		
		6.19	6.58		
		6.22			
T_j	$T_1 = 31.59$	$T_2 = 50.22$	$T_3 = 45.42$	$T_4 = 24.92$	$T = 152.15$
n_j	$n_1 = 5$	$n_2 = 8$	$n_3 = 7$	$n_4 = 4$	$N = 24$
Mean	6.318000	6.277500	6.488571	6.230000	6.339583

IIT KHARAGPUR

NPTEL ONLINE CERTIFICATION COURSES

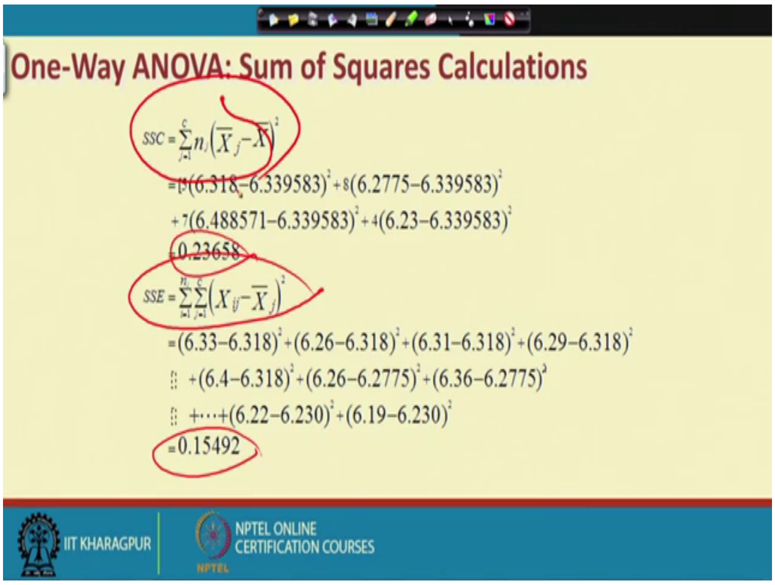
NPTEL

And in this case we have actually you know 4 samples and this is the sample 1, sample 2, sample 3, sample 4 and these are you know sample descriptions and this is the total of all these samples and by default mean of the sample will be 31.59 divided by 5 similarly 50 by 22 divided by 8 45.42 by 7 and so on like this.

So, accordingly mean is calculated and you can find out the variance also within this particular cell, but here actually variance should not be 0 in any point of time if variance is 0 for any particular sample then this particular sample should not be included in the particular process of investigation. So, now having this kind of and this is kind of descriptive statistics then we will go to the ANOVA structure and check whether there is a significant difference you know within the group and between the groups.

So, corresponding to this and the particular structure will be like this.

(Refer Slide Time: 22:19)



One-Way ANOVA: Sum of Squares Calculations

$$SSC = \sum_{j=1}^{c} n_j\left(\overline{X}_j - \overline{\overline{X}}\right)^2$$

$$= 10(6.318-6.339583)^2 + 8(6.2775-6.339583)^2$$

$$+ 7(6.488571-6.339583)^2 + 4(6.23-6.339583)^2$$

$$= 0.23658$$

$$SSE = \sum_{i=1}^{n_j}\sum_{j=1}^{c}\left(X_{ij} - \overline{X}_j\right)^2$$

$$= (6.33-6.318)^2 + (6.26-6.318)^2 + (6.31-6.318)^2 + (6.29-6.318)^2$$

$$+ (6.4-6.318)^2 + (6.26-6.2775)^2 + (6.36-6.2775)^2$$

$$+ \cdots + (6.22-6.230)^2 + (6.19-6.230)^2$$

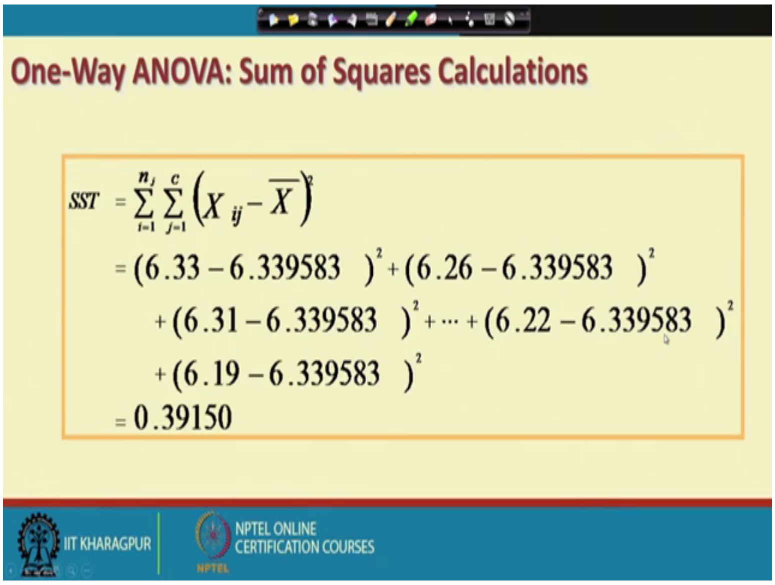
$$= 0.15492$$

So, first you know with the help of this particular formula and you can calculate the SSC and you can also calculate the SSE and once you get the SSC and SSE. So, by default you know sum of these 2 will be equal to called as you know SST that is the total sum of squares.

So, now, this is not actually difficult to refer to you know if you have a small kind of sample sorry sample numbers are you know small and then the sample observations are some you know small and in fact, if you actually you know excel spreadsheet or any kind of statistical software then just easily actually enter the data and you can have all these you know ANOVA items right.

So, these are all actually first time ANOVA item that is SST, SSC and SSE. So, once you get you know SST, SSC and SSE then with the help of degree of freedoms. So, you will find MSC and MSE. So, now, the ratio between the 2, with you know degree of freedoms. So, we will get the kind of F statistics and then that F statistic will compare with the critical value and accordingly we can get the inference and we will go for some kind of typical management decisions to need to be taken for this problem.

(Refer Slide Time: 23:44)



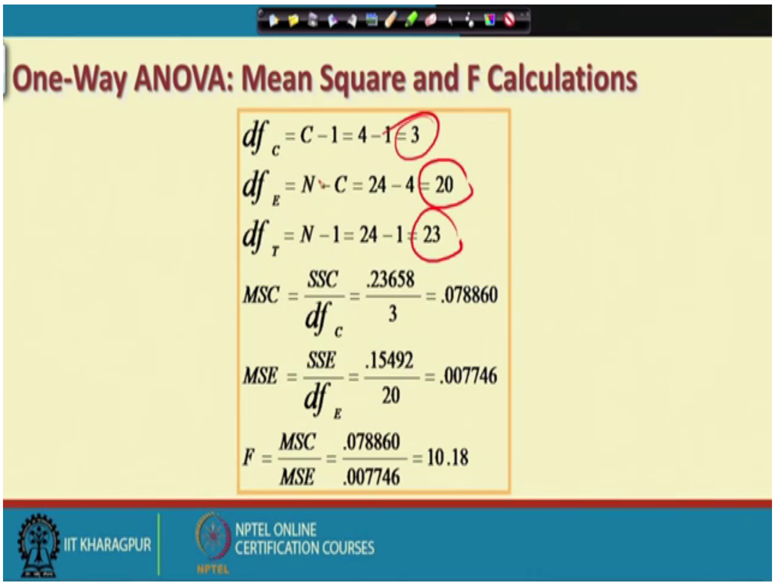
One-Way ANOVA: Sum of Squares Calculations

$$SST = \sum_{i=1}^{n_j}\sum_{j=1}^{c}\left(X_{ij}-\overline{X}\right)^2$$

$$= (6.33-6.339583)^2+(6.26-6.339583)^2$$
$$+(6.31-6.339583)^2+\cdots+(6.22-6.339583)^2$$
$$+(6.19-6.339583)^2$$
$$= 0.39150$$

And in this case in this case. So, this is how SST either you can calculate or you just add off it.

(Refer Slide Time: 23:48)



One-Way ANOVA: Mean Square and F Calculations

$$df_C = C-1 = 4-1 = 3$$
$$df_E = N-C = 24-4 = 20$$
$$df_T = N-1 = 24-1 = 23$$
$$MSC = \frac{SSC}{df_C} = \frac{.23658}{3} = .078860$$
$$MSE = \frac{SSE}{df_E} = \frac{.15492}{20} = .007746$$
$$F = \frac{MSC}{MSE} = \frac{.078860}{.007746} = 10.18$$

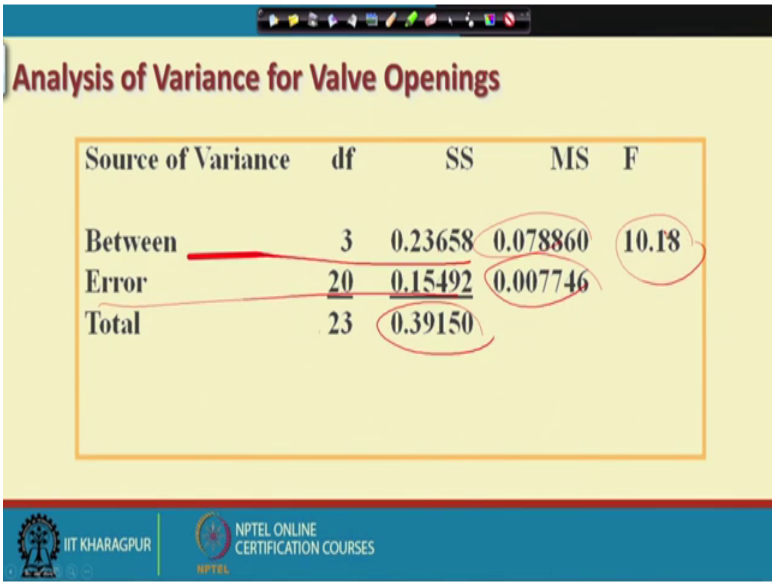
So, in order to know the details let us come here and this is what actually SSE which you have already calculated from this from this samples 2 3 6 5 8 and this is 1 4 9 2 and this is actually SST 2 3 9 1 5 0, but SST we have you know you know business, but our business is with respect to MSC and MSE. So, that is nothing, but SSC and SSE.

So, with respect to degree of freedom, you have to calculate. So, typically a you know you know SST degree of freedom is a N minus 1 and then the MSC degree of freedom is nothing, but actually C minus 1 that is actually column representations and then N minus C that is total samples minus columns will they give you the degree of freedom for the other part that is the MSE part. So, that is in this particular case in this particular case, this is for SST, this is for SSE and this is for SSC.

So, then accordingly the degree of freedom will be 3 for MSC and then for MSC it is actually 20 and then for SST, it is actually 23. So, accordingly, you can actually calculate it and then the 10.18 this is how the F calculated value. Then you go to the F critical and this is what actually F calculated and then you can go to the F criticals and then you have to take the decisions.

So, now corresponding to this you know problem, this is what exactly the ANOVA tables.

(Refer Slide Time: 25:22)



Analysis of Variance for Valve Openings

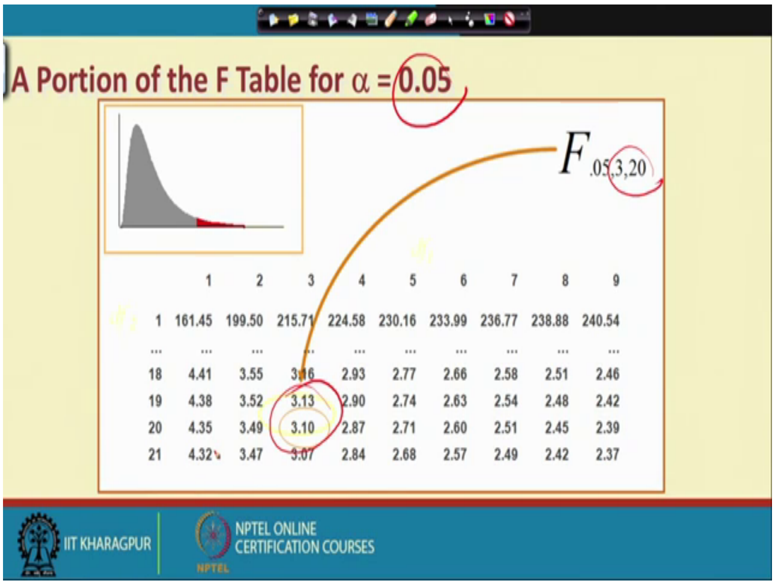
Source of Variance	df	SS	MS	F
Between	3	0.23658	0.078860	10.18
Error	20	0.15492	0.007746	
Total	23	0.39150		

IIT KHARAGPUR
NPTEL ONLINE CERTIFICATION COURSES

So, this is the between difference this is the between difference and this is the kind of within difference and these 2 will give you the kind of total and that is what SST is all about. So, now, this is mean and divide by degree of freedom you look at this divided by degree of freedom will get this one. So, this divided by this, divided by this degree of freedom will get this once.

So, now the ratio between the 2 will give you F statistic that is F calculated. Now, so to get to know the inference about this particular you know calculated which is you on the basis of sample information and then our fixing alpha and we have already degree of freedom on the basis of samples and we would like to know what is the critical value. So, the critical value for this particular you know structure is it like this here. So, fixing alpha equal to 0.05 and your degree of freedom 320, that means 320 here you go this way and this way then you will find here the kind of critical value.

(Refer Slide Time: 26:16)



A Portion of the F Table for α = 0.05

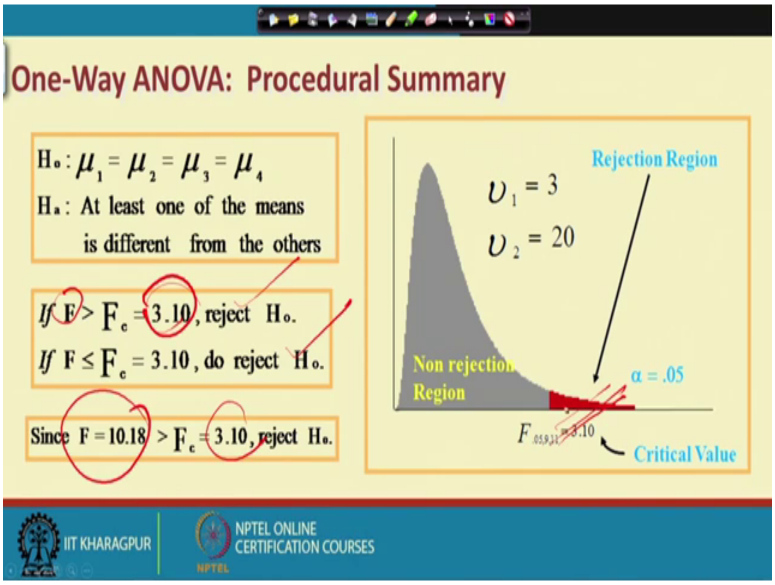
$F_{.05,3,20}$

	1	2	3	4	5	6	7	8	9
1	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54
...	...	...	...	...	...	...	...	...	...
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37

So, now on the basis of here you know F calculated and F criticals. So, you have to take the decision whether you have to accept the null hypothesis and whether to the knowledge; that means, the idea is a whether the population means are equal or population means are you know indifferent to each other, right.

So, in this contest, this is what actually F criticals. So, now, in the next slides , the typical you know inference will be 3.10 is the kind of range which you need to take a decisions whether your calculated value is over taking 3.10 or except you know under the 3 point you know a calculate critical value.

(Refer Slide Time: 26:57)



One-Way ANOVA: Procedural Summary

$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$

H_a: At least one of the means is different from the others

If $F > F_c = 3.10$, reject H_0.

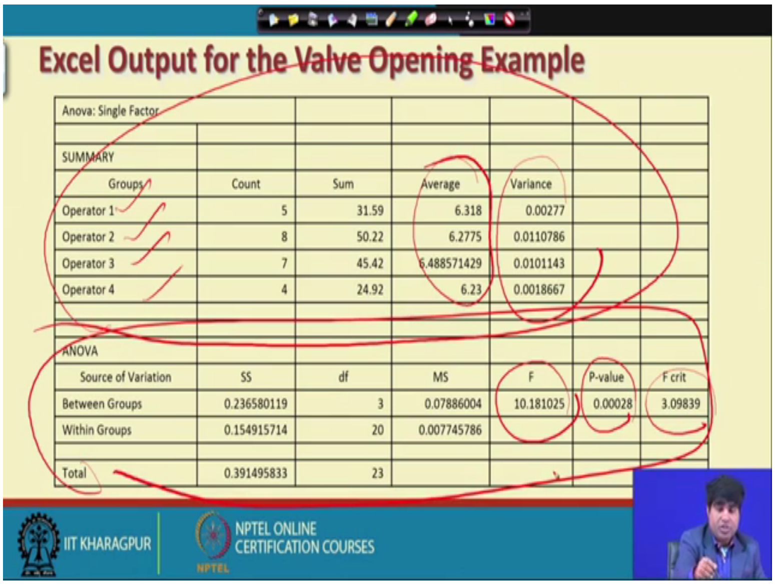
If $F \leq F_c = 3.10$, do reject H_0.

Since $F = 10.18 > F_c = 3.10$, reject H_0.

So, that means typically the whether you know you are you know on the basis of sample informations whether your critical value is overtaking the calculated value over taking the critical value or under the critical value. So, then accordingly you have to (Refer Time: 27:31). So, that means typically. So, this is actually critical value. So, this is what I have calculated. So, if it will be crossing this critical value, then you are rejecting true null hypothesis if not then you have to accept the true null hypothesis. So, in this case actually F calculated is coming 10.18 and then F critical is already given 3.10 since you have 10.1 it actually overtaking 3.10. So, that means we are here actually. So, that means we are rejecting the true null hypothesis and then the inference or the kind of conclusion is that mu 1, mu 2, mu 3, mu 4 are not same they are you know different and since they are you know different, that means there is a significant difference among these sample means. So, now, this will give you more interesting that you know which 2 pairs are you know actually problematic. So, and for that we may go for pairwise comparisons.

So, now, in this particular situations the conclusion is that you know there is a significance difference between the within the sample and between the samples. And accordingly, here we are you know rejecting the true null hypothesis and then the conclusion is that you know the mean sorry means of all these population are not saying they are actually different, alright.

(Refer Slide Time: 28:56)



Excel Output for the Valve Opening Example

Anova: Single Factor						
SUMMARY						
Groups	Count	Sum	Average	Variance		
Operator 1	5	31.59	6.318	0.00277		
Operator 2	8	50.22	6.2775	0.0110786		
Operator 3	7	45.42	6.488571429	0.0101143		
Operator 4	4	24.92	6.23	0.0018667		
ANOVA						
Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	0.236580119	3	0.07886004	10.181025	0.00028	3.09839
Within Groups	0.154915714	20	0.007745786			
Total	0.391495833	23				

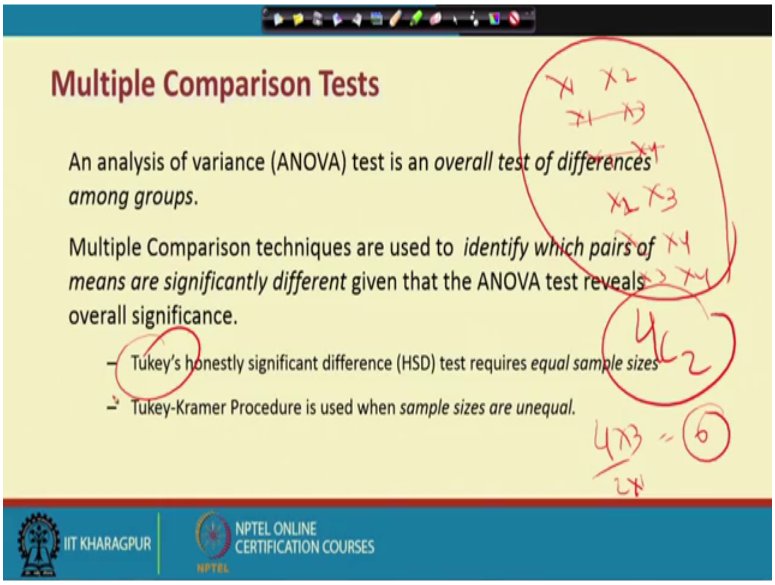
IIT KHARAGPUR

NPTEL ONLINE CERTIFICATION COURSES

So, this is what you know another way you can highlight this particular problem and the typical you know if you know this is my you know, 1-way ANOVA and this is what actually the kind of sample statistics and the corresponding ANOVA here is it is a sample statistic followed by mean of all these you know our samples 1 2 3 4 and corresponding variance. So, the variance itself showing that you know it is not a problems and it is under controls because you know if lower the variance then it will be more consistent of this particular series higher the variance this is a problematic and again if it is 0 then I still it is a problematic and since variance are not actually equal and not high. So, the problem is actually under controls and we are in a position to check now whether there is significant difference between their, you know the variance between the group and within the groups say. So, in this case, it is coming actually 10.18 and that is crossing it to layer criticals at the probability level of 5 percent.

And then the conclusion a conclusion is that you know the sample means are not actually same they are you know different. So, which sample is a you know having a high difference or low difference that will go that will take care by pairwise mean comparison again and for that we have a different approach or again and let us see how is this particular you know structure altogether.

(Refer Slide Time: 30:25)



Multiple Comparison Tests

An analysis of variance (ANOVA) test is an *overall test of differences among groups*.

Multiple Comparison techniques are used to *identify which pairs of means are significantly different* given that the ANOVA test reveals overall significance.

- Tukey's honestly significant difference (HSD) test requires *equal sample sizes*
- Tukey-Kramer Procedure is used when *sample sizes are unequal*.

x1 x2
x1 x3
x1 x4
x2 x3
x2 x4
x3 x4
$^4C_2$
$\frac{4 \times 3}{2 \times 1} = 6$
IIT KHARAGPUR
NPTEL ONLINE CERTIFICATION COURSES

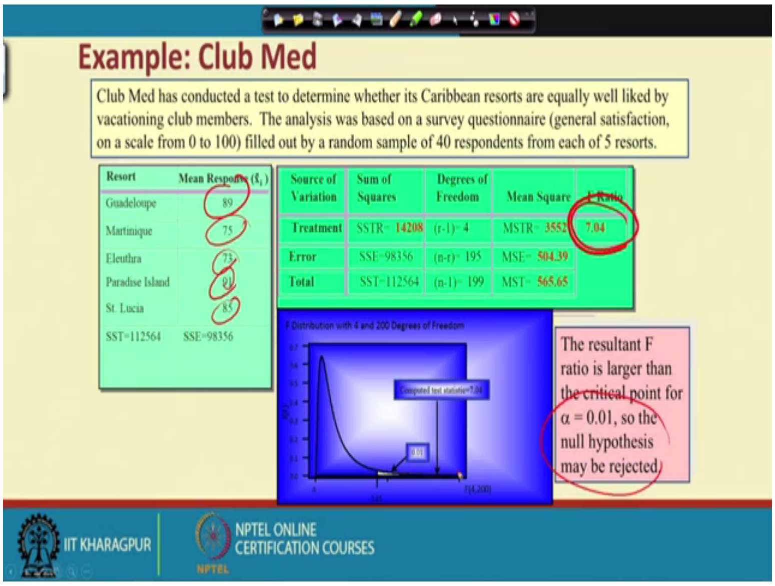
In the case of multiple comparison test, once ANOVA will give you some kind of significant result then; that means, it will give you another kind of route where we like to know which 2 pairs are actually a problematic. If they are not statistically if F is not showing start you know statistically significant then you may stop but if it is coming significant so that means one particular samples may be problematic.

Since here there are 4 sample you know case. So, or the pairwise comparison case will be a 4 c 2 times right. So, these are the, that means they typically 4 into 3 by 2 into 1, that means, technically 6 different you know cases you can find out and then you check what is the which 2 pairs are you know significantly different right. So that means, actually like you know x 1, x 2, x 1, x 3, x 1, x 4, x 1, x 1, x you know with 4 samples x 1 x 2, x 1 x 3, x 1 x 4, then x 2 x 3, then x 2 x 4 then finally, x 3 x 4. So, these are the following combinations you have to find and then you check which particular combination is again problem. So, then the management decision will be restricted to that particular difference only.

So, as a result since in this problem F is coming significant. So, it is the mandatory requirement that do we like to check which 2 pairs are again which particular pair or few pairs are statistically different. So, against for that we have to go for pairwise mean comparisons and in this case we use actually there are many methods are there, but we have to follow this particular method to case kind of pairwise mean comparisons and then we check which 2 pairs are you know statistically different so far as this particular problem is concerned.

So, accordingly, we have to move or to check you know the pairwise comparisons.

(Refer Slide Time: 32:34)



Example: Club Med

Club Med has conducted a test to determine whether its Caribbean resorts are equally well liked by vacationing club members. The analysis was based on a survey questionnaire (general satisfaction, on a scale from 0 to 100) filled out by a random sample of 40 respondents from each of 5 resorts.

Resort	Mean Response ($\bar{x}_i$)
Guadeloupe	89
Martinique	75
Eleuthra	73
Paradise Island	91
St. Lucia	85
SST=112564	SSE=98356

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F Ratio
Treatment	SSTR= 14208	(r-1)= 4	MSTR= 3552	7.04
Error	SSE=98356	(n-r)= 195	MSE= 504.39	
Total	SST=112564	(n-1)= 199	MST= 565.65	

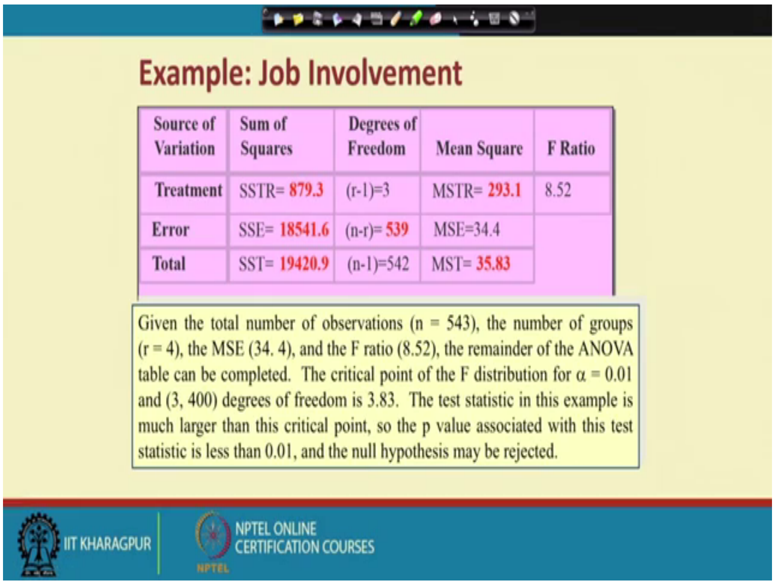
F Distribution with 4 and 200 Degrees of Freedom
Computed test statistic=7.04
0.01
F(4,200)

The resultant F ratio is larger than the critical point for $\alpha = 0.01$, so the null hypothesis may be rejected.

And here you see here is corresponding to this you know 5 different situations. So, here these are all actually these are all called as you know mean of this particular series and this mean of the individual series then we have a combined means right. So, then we like to check actually whether the mean difference among these individuals are statistically different or not.

But overall the problem is showing that you know they are you know statistically different for instance in this problem. So, F is coming in 7.04 and we are actually we can reject this hypothesis on the basis of the kind of F criticals and a fixing the kind of a critical values. So, we can see a you know what should be the kind of decision which you have take and to know which 2 pairs are you know statistically different to each other, alright.

(Refer Slide Time: 33:43)



Example: Job Involvement

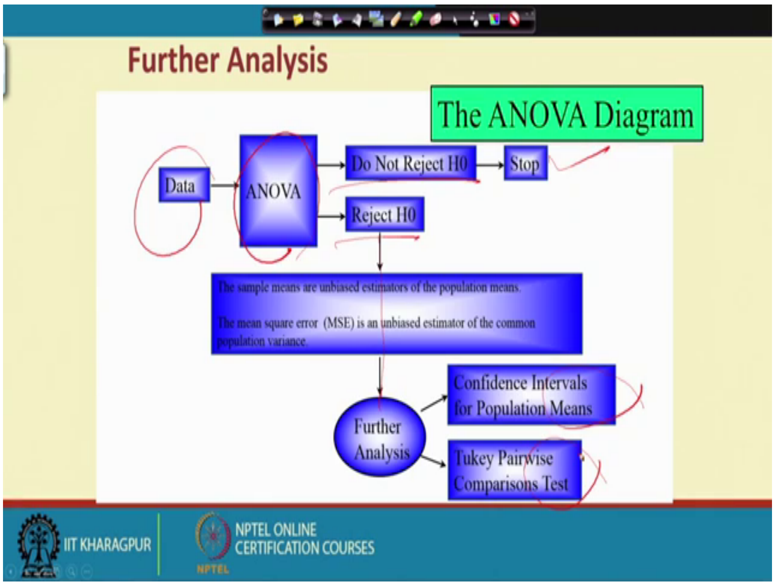
Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F Ratio
Treatment	SSTR= 879.3	(r-1)=3	MSTR= 293.1	8.52
Error	SSE= 18541.6	(n-r)= 539	MSE=34.4	
Total	SST= 19420.9	(n-1)=542	MST= 35.83	

Given the total number of observations (n = 543), the number of groups (r = 4), the MSE (34. 4), and the F ratio (8.52), the remainder of the ANOVA table can be completed. The critical point of the F distribution for α = 0.01 and (3, 400) degrees of freedom is 3.83. The test statistic in this example is much larger than this critical point, so the p value associated with this test statistic is less than 0.01, and the null hypothesis may be rejected.

IIT KHARAGPUR NPTEL ONLINE CERTIFICATION COURSES

So, this is the same things and you know it is a kind of the kind of difference between 2 sample means I am checking whether it is a statistically significant or not.

(Refer Slide Time: 33:53)



Further Analysis

The ANOVA Diagram

Data
ANOVA
Do Not Reject H0
Stop
Reject H0
Further Analysis
Confidence Intervals for Population Means
Tukey Pairwise Comparisons Test

The sample means are unbiased estimators of the population means.

The mean square error (MSE) is an unbiased estimator of the common population variance.

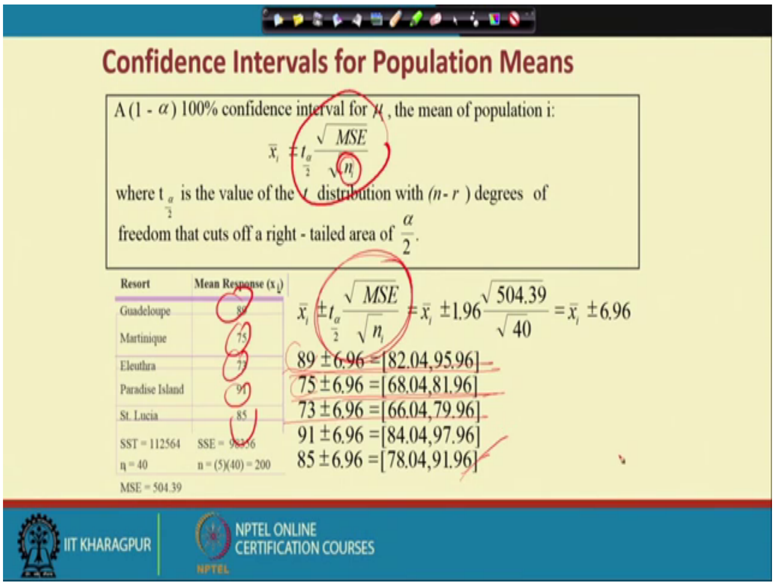
IIT KHARAGPUR NPTEL ONLINE CERTIFICATION COURSES

So, this is what the kind of. So, here we are actually to go for next level kind of inference and in that case the typical structure of ANOVA is like this. So, having the sample informations, we are using the ANOVA and then either you are rejecting H 0 or are not rejecting the H 0.

Do not reject H 0 means actually you have to stop, the problem may you know come to end there, but if you are rejecting the true null hypothesis then; obviously, the further issue will be starting and either you create a confidence interval for these population means and then we will check the pairwise mean comparisons.

So, like previously once you know once you test then accordingly you will see the confidence interval of this population mean or population mean difference and then we would like to check whether this you know on the pairs or. Obviously, if you have F is coming significant means any particular at least one particular pair will be statistical again different and this that particular pair will give you the typical management decision what should be the next you know level kind of inference or the kind of generic in your strategies right. So, accordingly we have to move.

(Refer Slide Time: 35:10)



Confidence Intervals for Population Means

A $(1-\alpha)$ 100% confidence interval for μ, the mean of population i:

$$\bar{x}_i \pm t_{\frac{\alpha}{2}} \frac{\sqrt{MSE}}{\sqrt{n_i}}$$

where $t_{\frac{\alpha}{2}}$ is the value of the t distribution with $(n-r)$ degrees of freedom that cuts off a right - tailed area of $\frac{\alpha}{2}$.

Resort	Mean Response (x_i)
Guadeloupe	89
Martinique	75
Eleuthra	73
Paradise Island	91
St. Lucia	85
SST = 112564	SSE = 98356
n_i = 40	n = (5)(40) = 200
MSE = 504.39	

$$\bar{x}_i \pm t_{\frac{\alpha}{2}} \frac{\sqrt{MSE}}{\sqrt{n_i}} = \bar{x}_i \pm 1.96 \frac{\sqrt{504.39}}{\sqrt{40}} = \bar{x}_i \pm 6.96$$

$89 \pm 6.96 = [82.04, 95.96]$

$75 \pm 6.96 = [68.04, 81.96]$

$73 \pm 6.96 = [66.04, 79.96]$

$91 \pm 6.96 = [84.04, 97.96]$

$85 \pm 6.96 = [78.04, 91.96]$

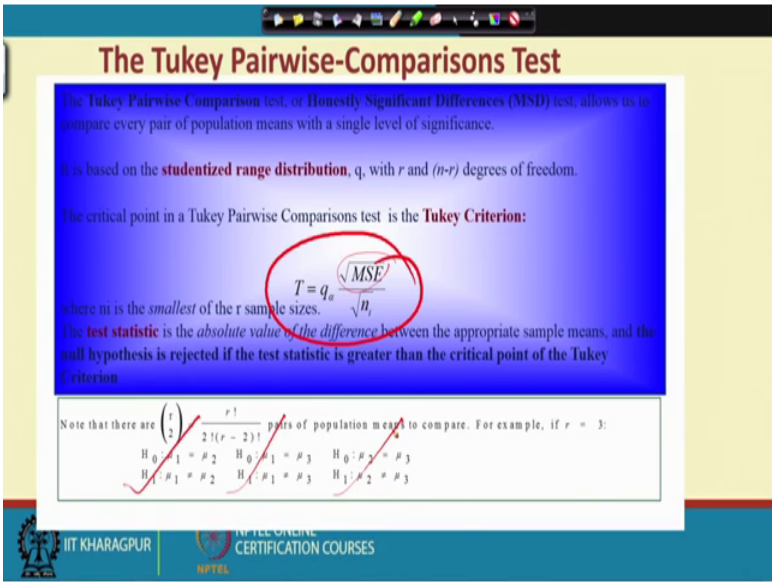
So, first of all actually we can create a confidence interval and corresponding to the previous problem corresponding to these previous problems. So, what we need to do.

So, this is what the problem. So, corresponding to this problem, now, the typical actually population interval will be for these are all individual samples then as we have already discussed for a particular you know population a sample means we can create a confidence interval and a having actually individual population means a sample means. So, you can go for you know population intervals right population interval.

So, for 89 the population interval will be like this, 75 the population interest, of course, this is adjusted with you know the kind of t statistics and because it is a pairwise you mean comparison and this is it and the best statistic which you apply here is the t statistic and this is how the t adjustment and a this is the sample specific because we have a different sample size and accordingly we have a different you know populace interval for different means right.

So, for particular means the confidence interval you know is different again for different means, in a confidence interval will be again actually different. So, likewise actually we have actually different kind of structure and this is the mean and this is what the confident interval for populations right. So, corresponding to this, we have to check actually the kind of significant difference and by using 2 case pairwise comparisons, this is what actually the test statistic which you have to calculate and that is depends upon you know you are in a sample size and mean square you know mean square MSE and that is actually within a group and then it will compare with you know individual test statistics. So, in the next slide I will give you the kind of exposures through which you have to take a kind of you know requires.

(Refer Slide Time: 36:48)



The Tukey Pairwise-Comparisons Test

The **Tukey Pairwise Comparison** test, or **Honestly Significant Differences (MSD)** test, allows us to compare every pair of population means with a single level of significance.

It is based on the **studentized range distribution**, q, with *r* and *(n-r)* degrees of freedom.

The critical point in a Tukey Pairwise Comparisons test is the **Tukey Criterion:**

$$T = q_\alpha \frac{\sqrt{MSE}}{\sqrt{n_i}}$$

where ni is the *smallest* of the r sample sizes.

The **test statistic** is the *absolute value of the difference* between the appropriate sample means, and **the null hypothesis is rejected if the test statistic is greater than the critical point of the Tukey Criterion**

Note that there are $\binom{r}{2} = \frac{r!}{2!(r-2)!}$ pairs of population means to compare. For example, if $r = 3$:

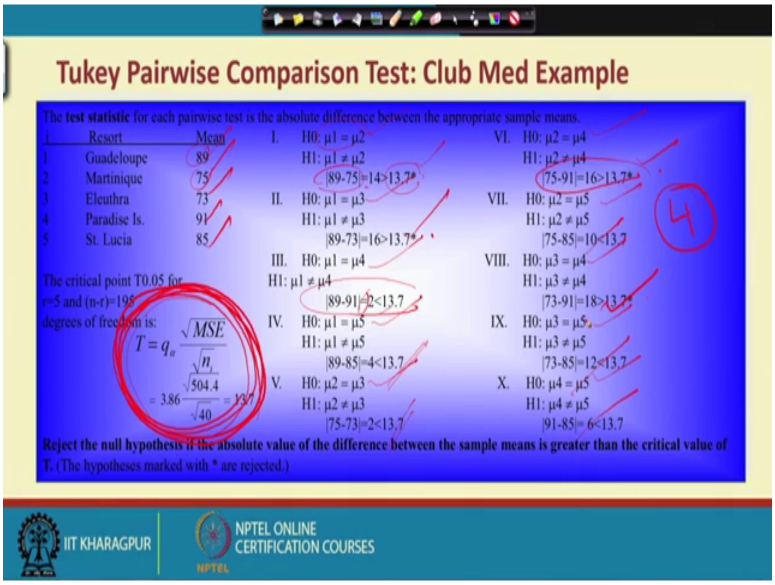
$H_0: \mu_1 = \mu_2$ $H_0: \mu_1 = \mu_3$ $H_0: \mu_2 = \mu_3$

$H_1: \mu_1 \neq \mu_2$ $H_1: \mu_1 \neq \mu_3$ $H_1: \mu_2 \neq \mu_3$

So, the null hypothesis will be the first you know the typical comparison will be like this and again we have a 6 different cases to test depending upon the particular you know case, you know situations right since we have actually 4 samples with you know 2 pairs then you know we have to go for c 2 if there are 4 5 different such cases then you will go for 5 c 2 pairs right.

So, accordingly in this problem, we will check how many cases we can actually test right.

(Refer Slide Time: 37:46)



Tukey Pairwise Comparison Test: Club Med Example

The **test statistic** for each pairwise test is the absolute difference between the appropriate sample means.

i	Resort	Mean
1	Guadeloupe	89
2	Martinique	75
3	Eleuthra	73
4	Paradise Is.	91
5	St. Lucia	85

The critical point T0.05 for r=5 and (n-r)=195 degrees of freedom is:

$$T = q_\alpha \frac{\sqrt{MSE}}{\sqrt{n_i}} = 3.86\frac{\sqrt{504.4}}{\sqrt{40}} = 13.7$$

I. H0: $\mu1 = \mu2$
H1: $\mu1 \neq \mu2$
$|89-75|=14>13.7$*

II. H0: $\mu1 = \mu3$
H1: $\mu1 \neq \mu3$
$|89-73|=16>13.7$*

III. H0: $\mu1 = \mu4$
H1: $\mu1 \neq \mu4$
$|89-91|=2<13.7$

IV. H0: $\mu1 = \mu5$
H1: $\mu1 \neq \mu5$
$|89-85|=4<13.7$

V. H0: $\mu2 = \mu3$
H1: $\mu2 \neq \mu3$
$|75-73|=2<13.7$

VI. H0: $\mu2 = \mu4$
H1: $\mu2 \neq \mu4$
$|75-91|=16>13.7$*

VII. H0: $\mu2 = \mu5$
H1: $\mu2 \neq \mu5$
$|75-85|=10<13.7$

VIII. H0: $\mu3 = \mu4$
H1: $\mu3 \neq \mu4$
$|73-91|=18>13.7$*

IX. H0: $\mu3 = \mu5$
H1: $\mu3 \neq \mu5$
$|73-85|=12<13.7$

X. H0: $\mu4 = \mu5$
H1: $\mu4 \neq \mu5$
$|91-85|= 6<13.7$

4

Reject the null hypothesis if the absolute value of the difference between the sample means is greater than the critical value of T. (The hypotheses marked with * are rejected.)

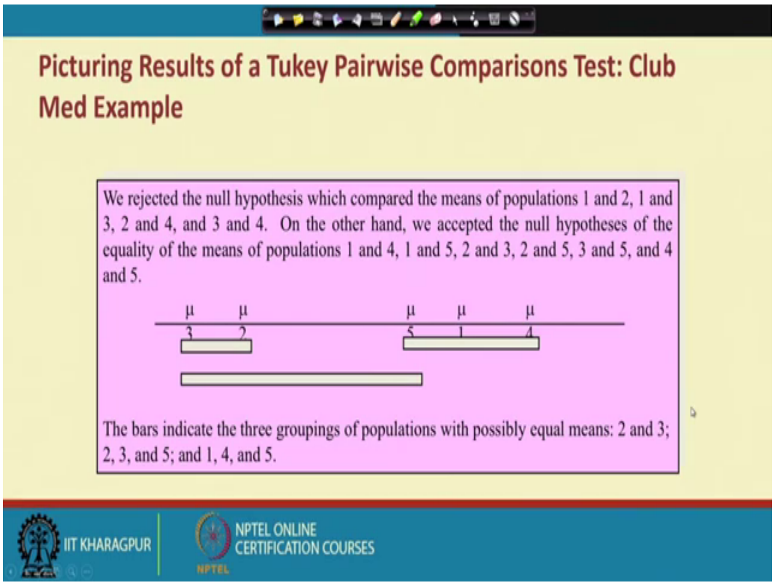
So, you see here is corresponding to this you know 5 samples corresponding 5 sample, sample 1, sample 2, sample 3, sample 4, sample 5 and this is what actually combined you know t statistics and that is the benchmark through which you have to check whether there is a significant difference. There is typically actually this is called as you know post hoc analysis right.

So, now the first case will be mu 1 mu 2 and mu 1 mu 3, mu 1 mu 4, mu 1 mu 5, similarly mu 2 mu 3, mu 2 mu 4, mu 2 mu 5, then mu 3 mu 4, mu 3 mu 5. So, these are you know possible cases with you know 5 different samples and mu 1 mu 2 equal to you know they are equal population mean are equal and with the help of sample means. So, here the sample means and with the sample means difference and that with respect to the benchmark to case t statistics. So, we will see actually whether this difference is actually over taking the kind of critical value.

So, in this case you know it is 14, you know so it is coming activate 13.7. So, that means it is a overtaking, so it is statistically different. And again this is also statistically different this is not statistically different because you know this value is lesser to the critical value this is again not statistically different. This is also not statistically different and this is again statistically a different. So, this is you know not decision this is against are still different, this is not, this is not. So, that means out of 10 different cases, 1 2 3 4. So, out of 10 cases, 4 different cases will find you know significant difference that is with respect to a sample 1 sample 2, sample 1 sample 3 and then sample 2 sample 4 and a finally, sample 3 and sample 4.

So, in other case there is no significant difference, that means, the management problem will be restricted to this samples now. Again, we have to check with you know where is the kind you know difference and how is this particular difference. So, the management problems accordingly or the management decision need to be taken care again as per the particular requirement.

(Refer Slide Time: 40:11)



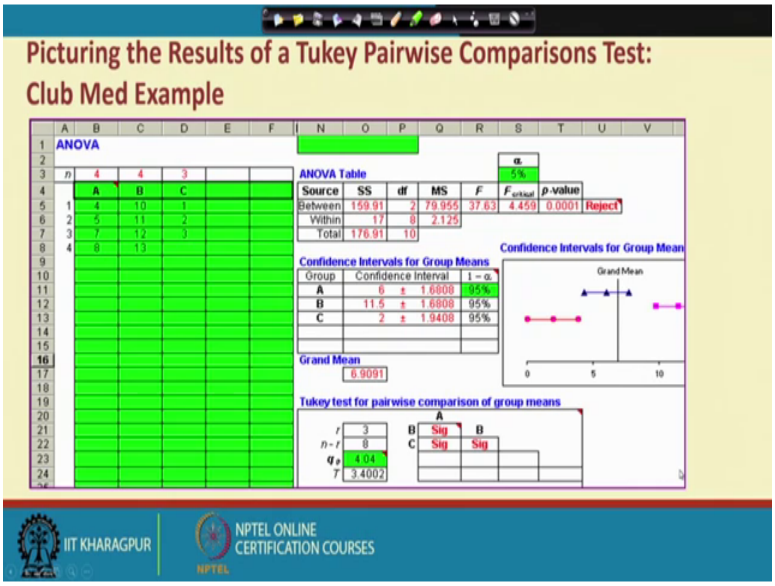
Picturing Results of a Tukey Pairwise Comparisons Test: Club Med Example

We rejected the null hypothesis which compared the means of populations 1 and 2, 1 and 3, 2 and 4, and 3 and 4. On the other hand, we accepted the null hypotheses of the equality of the means of populations 1 and 4, 1 and 5, 2 and 3, 2 and 5, 3 and 5, and 4 and 5.

The bars indicate the three groupings of populations with possibly equal means: 2 and 3; 2, 3, and 5; and 1, 4, and 5.

So, this is how the typical actually a graphical you know structuring a regarding the difference between the 2 different you know pairwise comparisons. So, with the help of 5 different situation earlier, 3 situations or 3 different occasions we are actually finding the difference or statistically different and in other cases they are not statistically different. So, where it is not statistically different, the problem will not be so serious, but where there is a statistical difference then the inference will be different and according to you know take into considerations and we analyze the problem as per the kind of management requirement.

(Refer Slide Time: 40:38)



Picturing the Results of a Tukey Pairwise Comparisons Test: Club Med Example

ANOVA

n	4	4	3	
	A	B	C	
1	4	10	1	
2	5	11	2	
3	7	12	3	
4	8	13		

ANOVA Table

α: 5%

Source	SS	df	MS	*F*	$F_{critical}$	*p*-value	
Between	159.91	2	79.955	37.63	4.459	0.0001	Reject
Within	17	8	2.125				
Total	176.91	10					

Confidence Intervals for Group Means

Group	Confidence Interval			1 − α
A	6	±	1.6808	95%
B	11.5	±	1.6808	95%
C	2	±	1.9408	95%

Grand Mean

6.9091

Confidence Intervals for Group Mean

Tukey test for pairwise comparison of group means

r	3
n − *r*	8
q_α	4.04
T	3.4002

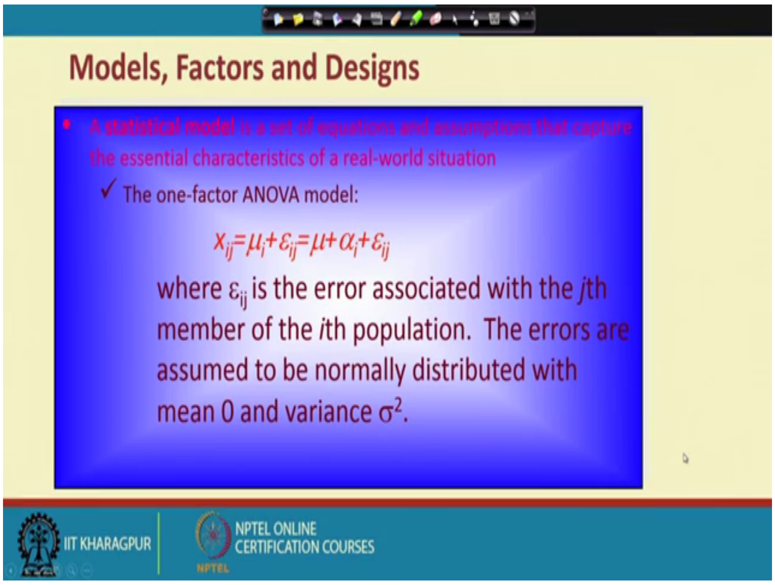
	A	B
B	Sig	
C	Sig	Sig

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

And this is another kind of kind of comparisons you can go to the excel sheet and then we will do some kind of comparative statistics and then you see. First you go for the kind of combined a kind of test that is through F statistics and if F is coming statistically significant then you have to see which particular pairs are actually reflecting.

If F is not coming significant then the next level sorts or next level inference is not required, but if F is coming statistically significant that is; that means, the situation where you know F calculated will about take the critical then you will find which pairs are you know again coming into the pictures; for that means, statistically different. And then you to you have to think how or how it can be actually materialized as per the kind of management need or management requirement.

(Refer Slide Time: 41:47)



Models, Factors and Designs

- A **statistical model** is a set of equations and assumptions that capture the essential characteristics of a real-world situation
  - ✓ The one-factor ANOVA model:

$$X_{ij}=\mu_i+\varepsilon_{ij}=\mu+\alpha_i+\varepsilon_{ij}$$

where ε_{ij} is the error associated with the *j*th member of the *i*th population. The errors are assumed to be normally distributed with mean 0 and variance σ^2.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

This is what actually the kind of you know, these are all again you know advanced level of kind of things like you know models factors and a designs typical design of experiments. So, that means the, so far as you know scope of ANOVA is concerned it has you know it has a kind of multiple impact or multiple use or having enough scope for integrating with the different problems and the same structure can be extended with you know 2-way ANOVA, 3 way ANOVA and kind of MANOVA, right.

So, in the mean times you know we will stop here and in the next lectures we will extend this ANOVA with you know 2-way contest and 3 way contest. And we will check, what is the kind of structure? And the kind of outcome and what should be the kind of typical management decision corresponding to multivariate case in that to in the case of 2-way ANOVA, 3-way ANOVA and MANOVA. With this we will stop here.

Thank you very much, Have a nice day.