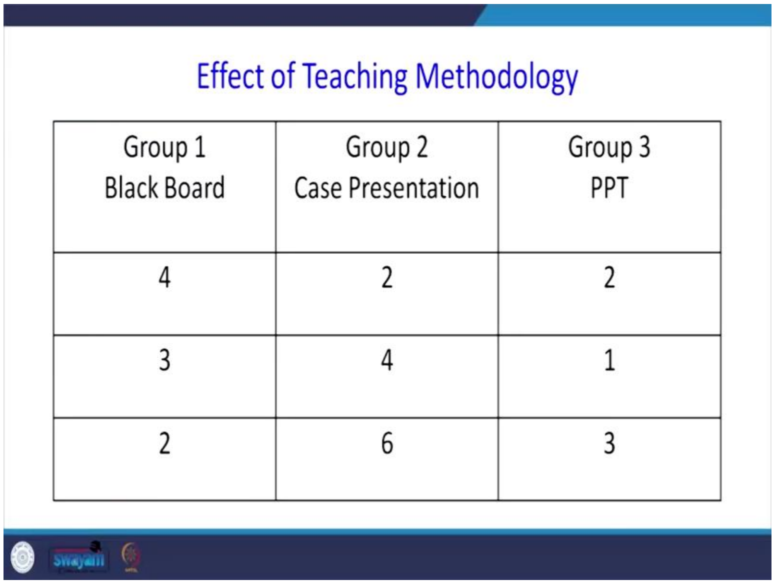
Data Analytics with Python
Prof. Ramesh Anbanandam
Department of Management Studies
Indian Institute of Technology – Roorkee

Lecture – 24
ANOVA- II

Dear students in the previous class I have explained the concept behind analysis of variance. In this class with the help of Python will solve that problem because previously in the previous class we have solved it to manually. Now we will use the help of Python will solve that problem this was the problem which was given.

(Refer Slide Time: 00:47)

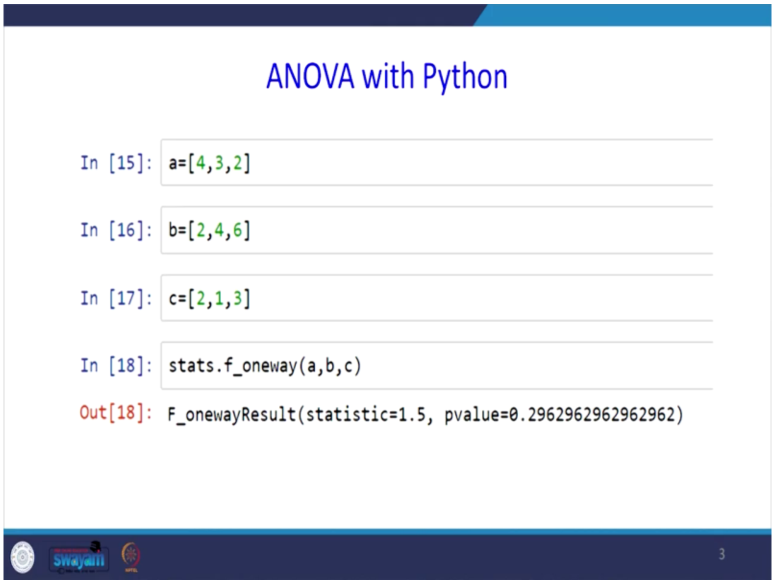


Effect of Teaching Methodology

Group 1 Black Board	Group 2 Case Presentation	Group 3 PPT
4	2	2
3	4	1
2	6	3

What is the problem is there are 3 different teaching methodology, we have to say which teaching methodology is more influencing on the student performance.

(Refer Slide Time: 01:01)

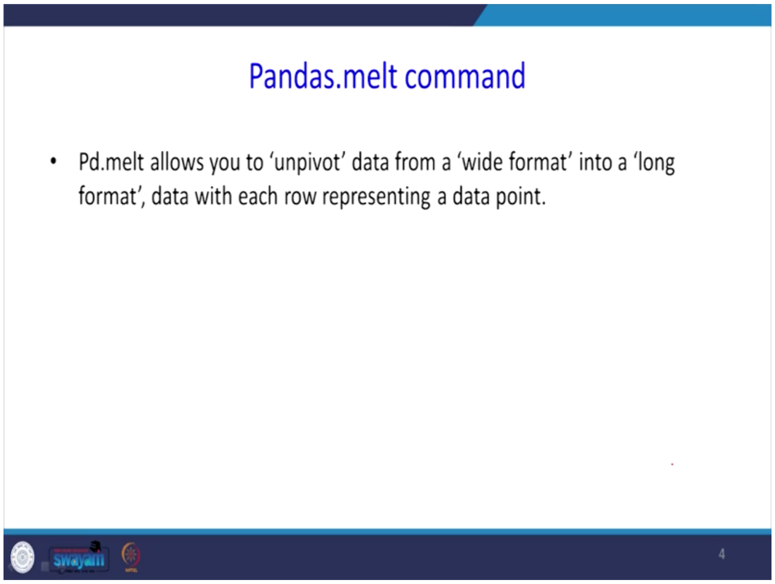


ANOVA with Python

```
In [15]: a=[4,3,2]

In [16]: b=[2,4,6]

In [17]: c=[2,1,3]

In [18]: stats.f_oneway(a,b,c)
Out[18]: F_onewayResult(statistic=1.5, pvalue=0.2962962962962962)
```

In Python so I have taken a is an array a 4, 3, 2 b 2, 4, 6 c 2, 1, 3 if you type stats dot F underscore one way just you call abc you run that one you will get this was were valuing this was for F that is a calculated F value, this is a p value. Suppose if the Alpha equal to 5 % it is more than 5 % so we have to accept a null hypothesis that is what our previous result also.

(Refer Slide Time: 01:32)

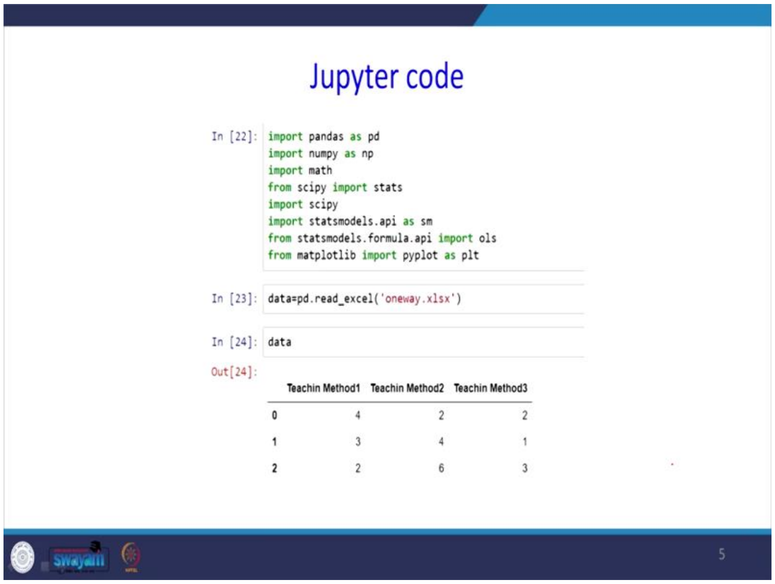


Pandas.melt command

- Pd.melt allows you to 'unpivot' data from a 'wide format' into a 'long format', data with each row representing a data point.

Now we will use that another command that that is pandas stat to melt command we will see the purpose of this command for doing ANOVA pd dot melt allows you to unpivot data from a wide format into long format that is data with each row representing a data point.

(Refer Slide Time: 01:51)



Jupyter code

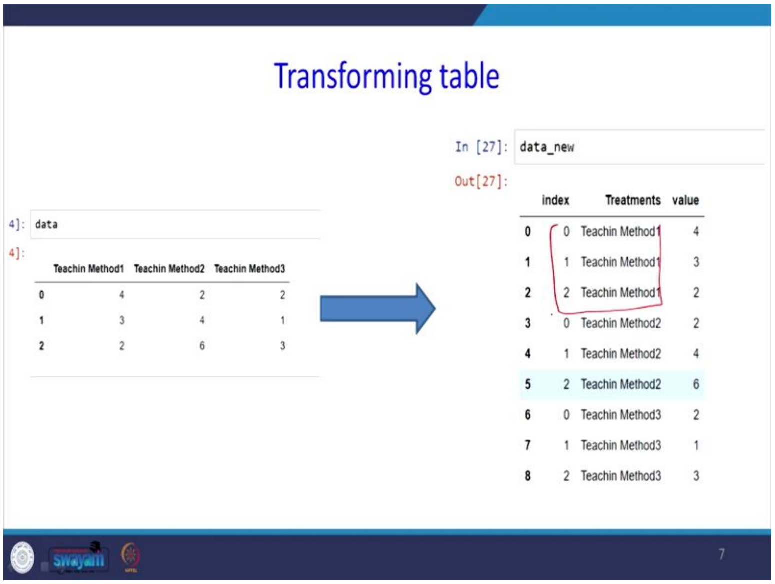
```
In [22]: import pandas as pd
         import numpy as np
         import math
         from scipy import stats
         import scipy
         import statsmodels.api as sm
         from statsmodels.formula.api import ols
         from matplotlib import pyplot as plt

In [23]: data=pd.read_excel('oneway.xlsx')

In [24]: data
Out[24]:
```

	Teachin Method1	Teachin Method2	Teachin Method3
0	4	2	2
1	3	4	1
2	2	6	3

So, for that purpose input pondas dot pd import numpy is np input math from scipy import stats import scipy import stats model dot api as sm from stats models dot formula dot api import wireless from matplotlib import pi plot as plt, so first we will load the data the data I am going to save the data the given time in the excel file are going to save in the object called data. So, data equal to PD dot read underscore Excel one way dot xlsx.

So, I have loaded when I run this data now the data is appearing column 1 column 2 column 3 so 0 1 2 that is your index.

(Refer Slide Time: 02:42)



Transforming table

```
4]: data
4]:
```

	Teachin Method1	Teachin Method2	Teachin Method3
0	4	2	2
1	3	4	1
2	2	6	3

```
In [27]: data_new
Out[27]:
```

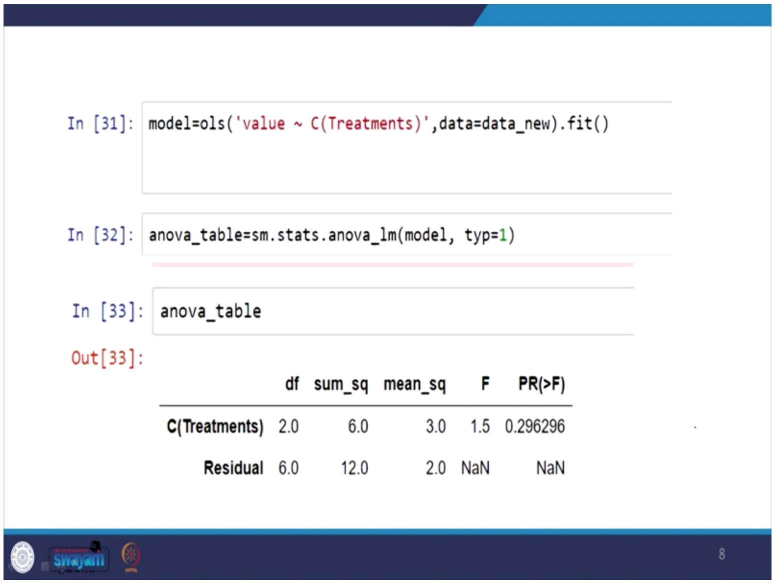
	index	Treatments	value
0	0	Teachin Method1	4
1	1	Teachin Method1	3
2	2	Teachin Method1	2
3	0	Teachin Method2	2
4	1	Teachin Method2	4
5	2	Teachin Method2	6
6	0	Teachin Method3	2
7	1	Teachin Method3	1
8	2	Teachin Method3	3

Next what you have to do for running the data I need to have the data in this format what is that format is so T.M. 1e teaching methodology one teaching methodology if your teaching methodology one there may be some numbers. In teaching methodology 2 there may be some numbers and teaching methodology 3 there may be some numbers. So, this says your treatment so the next one says the value suppose I want to have the data in this format.

For that you have to use the following command that is data underscore new I am going to call it is that way pd dot melt data dot reset underscore index parentheses, ID underscore VARS equal to in the square bracket index, value underscore VRS in square bracket teachin method 1 teachin method 2 teachin method 3 data underscore new dot column equal to index, treatment, value. So, this is the this is the syntax for using the melt function.

So, if you run this data underscore new will get this kind of odd for do you see previously the data was 0 1 2 format now what we are saying all teaching methodology one it is grouped in this way this is group 1 this is your group 2 this is group 3. Now only one there was only 2 column one is a one is treatment another one is value. now after getting this data into this format for converting this purpose the melt command is used.

(Refer Slide Time: 04:22)

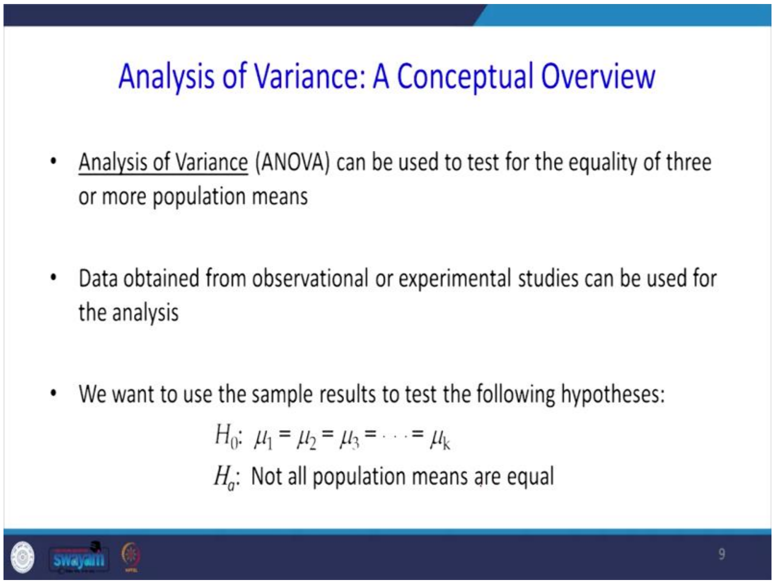


```
In [31]: model=ols('value ~ C(Treatments)',data=data_new).fit()

In [32]: anova_table=sm.stats.anova_lm(model, typ=1)

In [33]: anova_table
Out[33]:
                 df  sum_sq  mean_sq    F    PR(>F)
C(Treatments)   2.0     6.0      3.0  1.5  0.296296
     Residual   6.0    12.0      2.0  NaN       NaN
```

So model equal to wireless, wireless is ordinary least square method in in quote value tilde C treatment, data equal to delta underscore new fit. Then if you write ANOVA underscore table equal to s m dot stats data on over underscore lm model type 1 when you run this will get the ANOVA table this represents degrees of freedom for treatment because there was a 3 column treatment is 2 whereas dual is 6 because for a column 1 there are 3 elements so 3 – 1, 2 degrees of freedom.

Similarly for column 2 also another 2 degrees of freedom for column 3 also another 2 degrees of freedom totally 6 degrees of freedom, so some squared is 6 here some of square is 12. So, mean some of square is 6 divided by 2 that is 6 this is 12 divided by 6 that is 2, so F value is 3 divided by 2 this was the p value this also we got it previously to help of; when you do it manually we compared this 1.5 and we got the same value this is with the help of Python we are also getting the same result.

(Refer Slide Time: 05:31)



Analysis of Variance: A Conceptual Overview

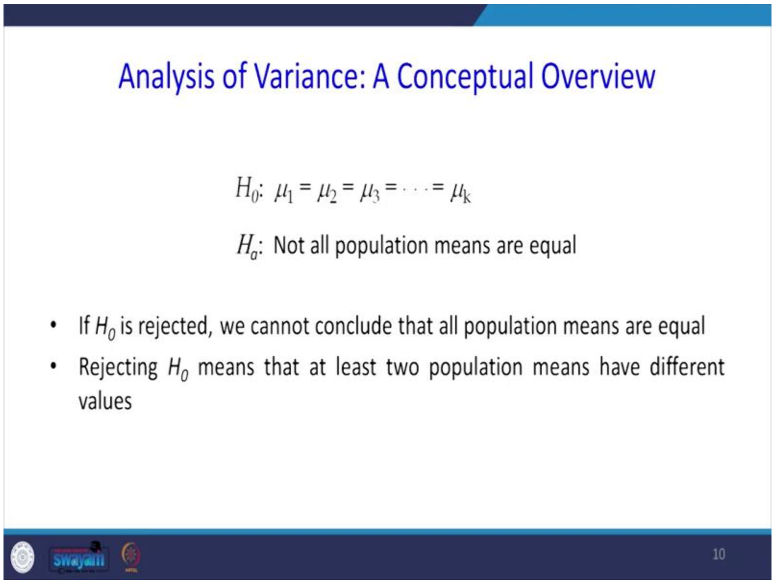
- Analysis of Variance (ANOVA) can be used to test for the equality of three or more population means
- Data obtained from observational or experimental studies can be used for the analysis
- We want to use the sample results to test the following hypotheses:

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

H_a: Not all population means are equal

Now we will go to the formal definition of ANOVA venire conceptual overview analysis of variance can be used to test the Equality of 3 or more population means. Data obtained from observational or experimental studies can be used for this analysis. We want to use the sampled result to test the following hypothesis in ANOVA what does the hypothesis H 0 equal to mu 1 equal to mu 2 equal to mu 3 it may be n number of columns.

(Refer Slide Time: 06:17)



Analysis of Variance: A Conceptual Overview

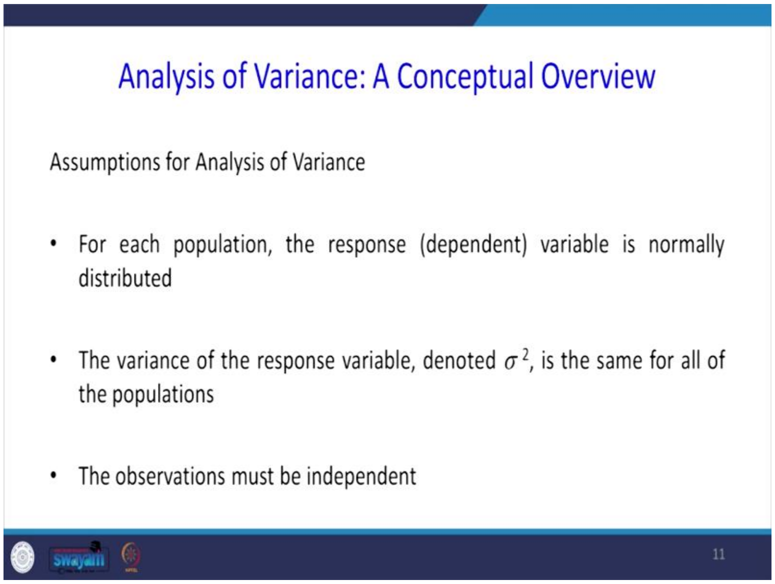
H_0: $\mu_1 = \mu_2 = \mu_3 = \cdots = \mu_k$

H_a: Not all population means are equal

- If H_0 is rejected, we cannot conclude that all population means are equal
- Rejecting H_0 means that at least two population means have different values

Alternative hypothesis not all population means are equal H 0 equal to mu 1 equal to mu 2 equal to mu 3 H a is not all population means are equal if it is not is rejected we cannot conclude that all population means are equal so when you reject it that means there are some unusual means. So, rejecting H 0 means that at least to 2 population means how different values.

(Refer Slide Time: 06:36)



Analysis of Variance: A Conceptual Overview

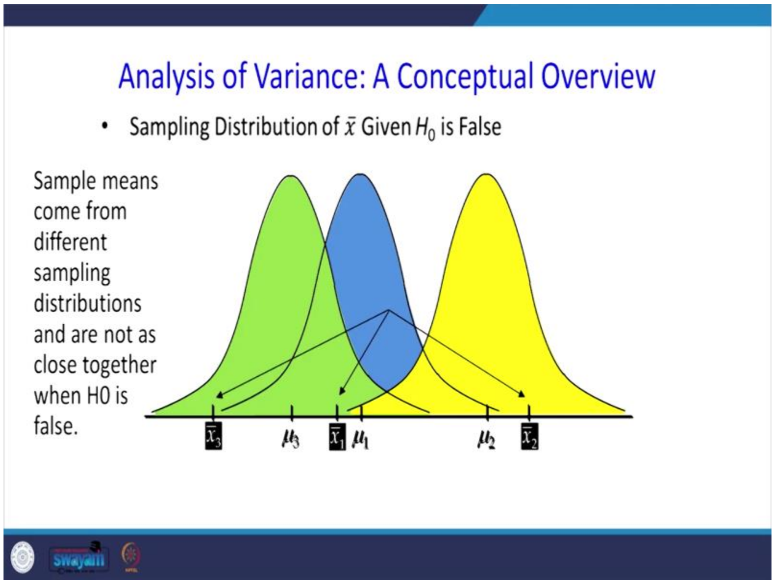
Assumptions for Analysis of Variance

- For each population, the response (dependent) variable is normally distributed
- The variance of the response variable, denoted σ^2, is the same for all of the populations
- The observations must be independent

What are the Assumption for analysis of variance. For each population the response dependent variable is normally distributed. In our previous example the performance of the student is the dependent variable the independent variable is teaching methodology. The variance of the response variable denoted by Sigma square is the same for all populations. Why this assumption is required? when you are comparing more than 2 groups the basic assumption is that the variance of that group should be same.

This concept we have explained when we are conducting 2 sample tests and the observation must be independent.

(Refer Slide Time: 07:16)



Analysis of Variance: A Conceptual Overview

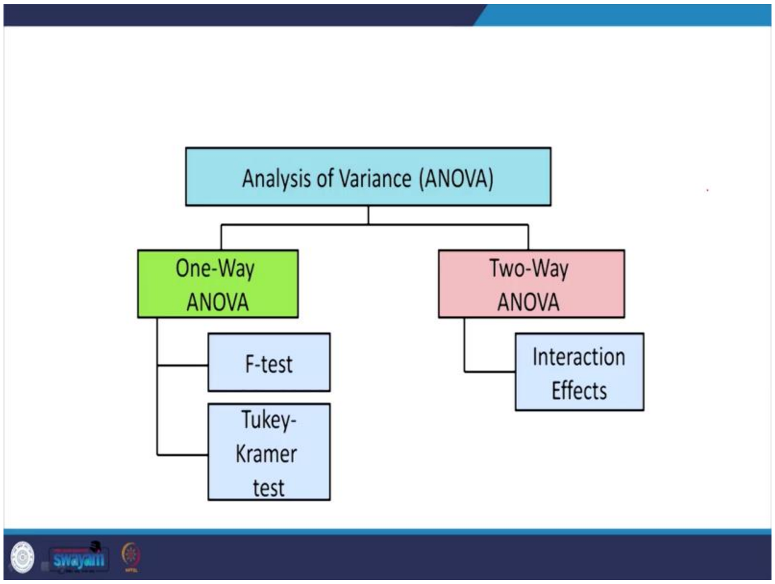
- Sampling Distribution of $\bar{x}$ Given H_0 is False

Sample means come from different sampling distributions and are not as close together when H0 is false.

Look at this normal distribution this is the sampling distribution of x-bar given null hypothesis is a true sample means are close together because there is only one sampling distribution when H 0 is true. Look at this normal distribution there are 3 normal distribution here the sampling distribution of X-bar given H 0 is false what is H 0 is false what is H 0 here H 0 equal to mu 1 equal to mu 2 equal to mu 3 if it is a false what will happen it would not be from same population it will be from different population.

So, the sample means come from different sampling distribution and are not as close together when H 0 is false.

(Refer Slide Time: 08:05)



Analysis of Variance (ANOVA)

One-Way ANOVA

F-test

Tukey-Kramer test

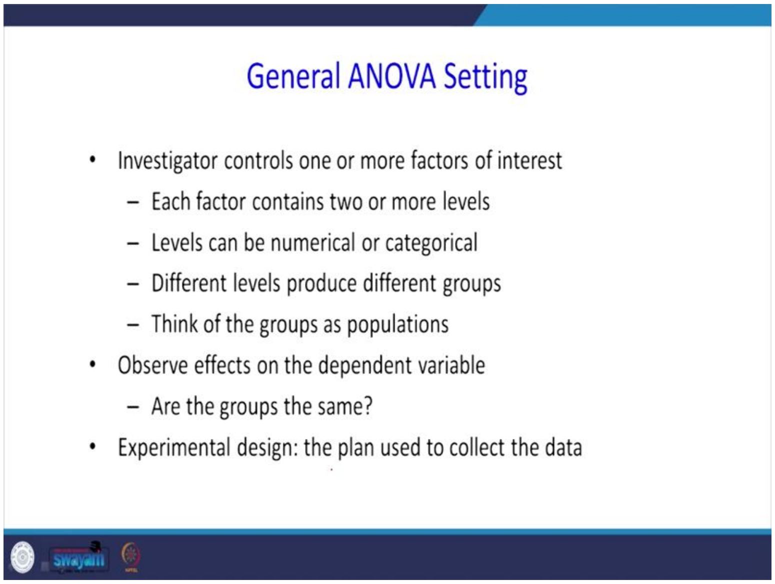
Two-Way ANOVA

Interaction Effects

In analysis of variance we can classify one is 1 way ANOVA another one is 2 way ANOVA there is one more thing in between that is colder R B D randomized block design will see when we will go for randomized block design. In one-way ANOVA we are going to do the F test the F test will help you to decide whether the null hypothesis accepted or rejected when you reject the null hypothesis then Tukey Kramer test will help you which 2 pairs are equal or which 2 pairs are not equal.

Then the this side is a 2-way ANOVA then we will go for interaction effects that we will see in coming classes. In this class we will see how to do the F test how will you do the 2 Tukey Kramer test.

(Refer Slide Time: 08:49)



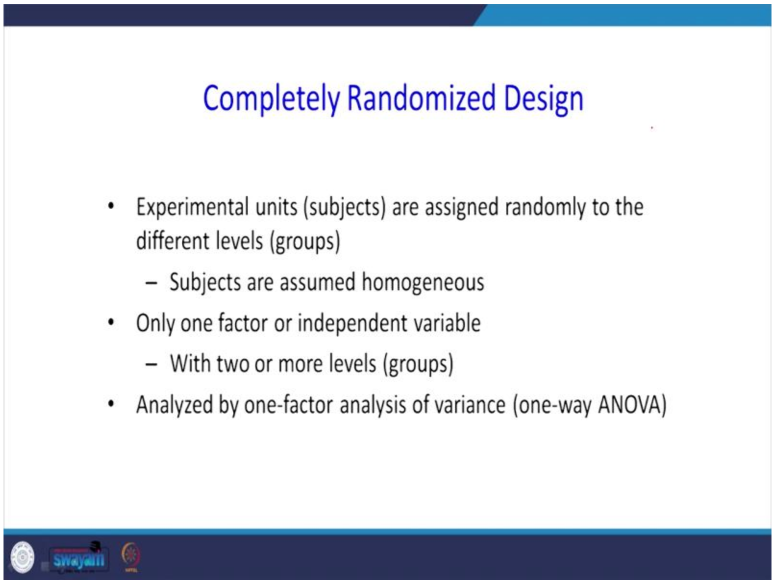
General ANOVA Setting

- Investigator controls one or more factors of interest
  - Each factor contains two or more levels
  - Levels can be numerical or categorical
  - Different levels produce different groups
  - Think of the groups as populations
- Observe effects on the dependent variable
  - Are the groups the same?
- Experimental design: the plan used to collect the data

What is the general ANOVA setting investigator controls one or more factors of interest in our previous example the teaching methodology is the factor. So, each factor contains 2 or more level in our case you see suppose of the pressure is the one parameter may be high or low that is level. So, high is one level low is another level, level can be numerical or categorical. Here the example is it is a categorical he need not be categorical it may be a continuous variable also.

So, different levels produce different groups think of the groups as population we can say each groups can be we can consider as the population. Observe effect on the dependent variable so what we are going to doing otherwise what is the effect of this treatment on the dependent variable. Next we will see experimental design the plan used to collect the data only the external design will have a plan to collect the data. And will see the effect of this data on the how this treatment is influencing the data.

(Refer Slide Time: 09:59)



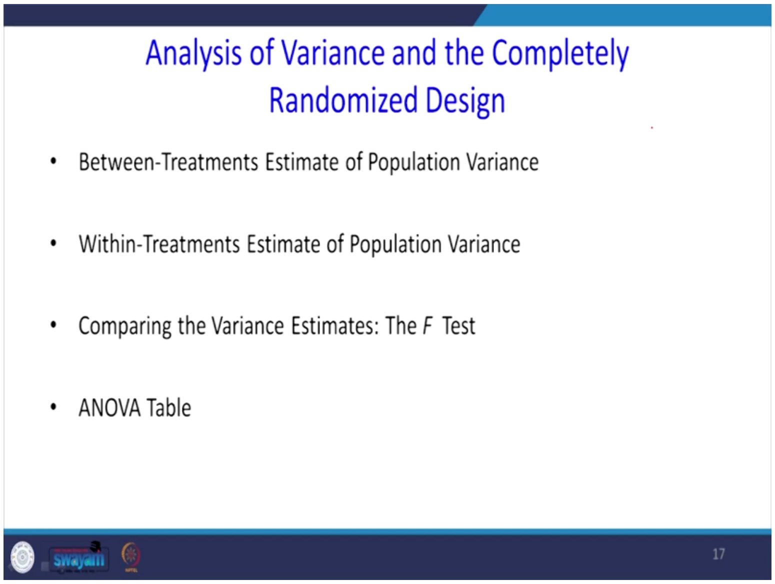
Completely Randomized Design

- Experimental units (subjects) are assigned randomly to the different levels (groups)
  - Subjects are assumed homogeneous
- Only one factor or independent variable
  - With two or more levels (groups)
- Analyzed by one-factor analysis of variance (one-way ANOVA)

The first one method called completely randomized design in our previous example the students are allocated to 3 groups randomly that is an example of you were completely randomized design. There is no bias because what will happen there suppose if you consider the student IQ level then you are allocating that is stored in 2 different category of classes then that is not called biased method. So, what is happening here the experimental units are assigned randomly to the different levels so subjects are assumed homogeneous.

Only one factor or one independent variable that is called one way ANOVA because here teaching methodology the independent variable the student performance that is the marks is the dependent variable. There also we can have 2 or more levels if you are analyzing one factor analysis of variance it is called one way one way ANOVA. If there are 2 independent variable that is a 2 way ANOVA.

(Refer Slide Time: 10:57)

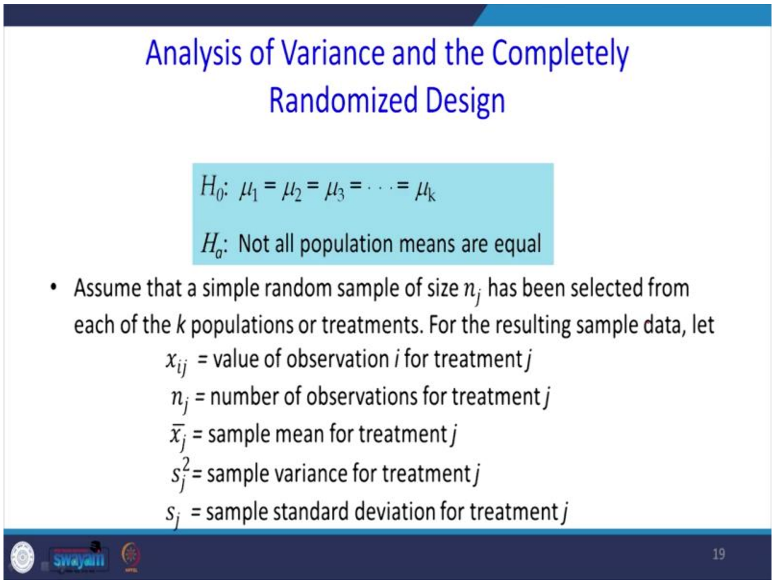


Analysis of Variance and the Completely Randomized Design

- Between-Treatments Estimate of Population Variance
- Within-Treatments Estimate of Population Variance
- Comparing the Variance Estimates: The F Test
- ANOVA Table

So, what we are doing the basic concept behind is we are finding the variance due to between the treatment and variance between the treatments. So, that will go to your numerator that is nothing but every SS treatment that is why I wrote SSB. When you divide by degrees of freedom this is MSB divided by the SSE divided by degrees of freedom that is variance within the treatment this is nothing but you are MSE. So, we will find variance between the treatment then variance within the treatment then we will go for F test then I will explain what is this ANOVA table.

(Refer Slide Time: 11:44)



Analysis of Variance and the Completely Randomized Design

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

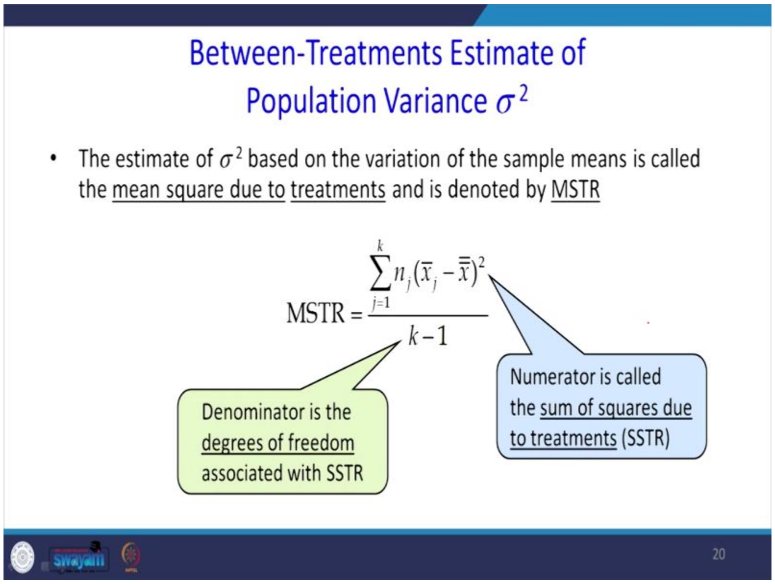
H_a: Not all population means are equal

- Assume that a simple random sample of size n_j has been selected from each of the k populations or treatments. For the resulting sample data, let

x_{ij} = value of observation i for treatment j
n_j = number of observations for treatment j
$\bar{x}_j$ = sample mean for treatment j
s_j^2 = sample variance for treatment j
s_j = sample standard deviation for treatment j

So, what is a null hypothesis for the CRD completely randomized design mu1 equal to mu2 mu3 equal to mu k not all population means are equal. Here assume that a simple random sample of nj has been selected from each of the k populations or treatment there are k treatment in our previous example there was a 3 treatment 3 factors 3 factors means 3 levels for the resulting sample data let X ij value of observation i for treatment j nj is number of observation for treatment j xj is sample mean for treatment say s j square sample variance for treatment j s j is the sample standard deviation for treatment j

(Refer Slide Time: 12:32)



Between-Treatments Estimate of Population Variance σ^2

- The estimate of σ^2 based on the variation of the sample means is called the mean square due to treatments and is denoted by MSTR

$$\text{MSTR} = \frac{\sum_{j=1}^{k} n_j(\bar{x}_j - \bar{\bar{x}})^2}{k-1}$$

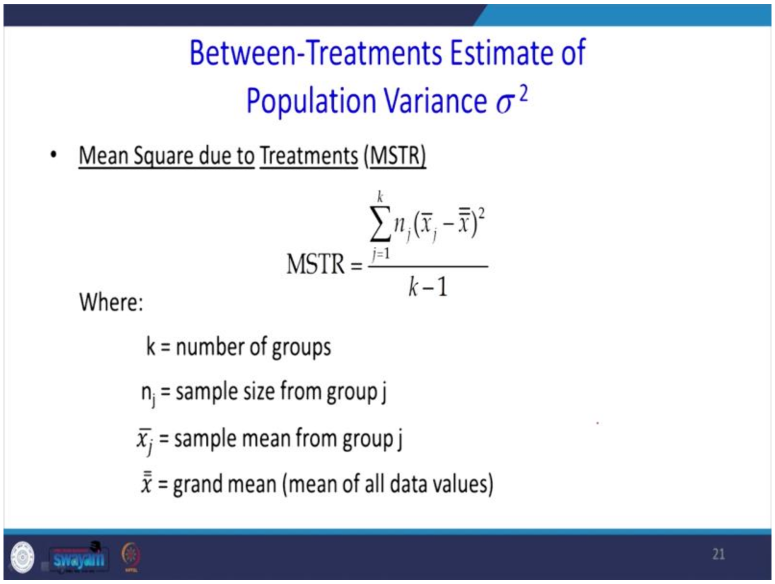
Denominator is the degrees of freedom associated with SSTR

Numerator is called the sum of squares due to treatments (SSTR)

First we will find out between treatment estimation of population variance Sigma square. The estimate of Sigma square based on the variation of the sample mean is called mean square due to treatment that is denoted by MSTR in our example previously we have used we can have MSB that is mean square due to between columns. So, how we are finding is this MSTR is nothing but n j number of elements in column j xj that column j mean minus the overall mean whole square divided by k – 1, k is the number of columns here 3 – 1.

So the denominator is the degrees of freedom associated with sum of square treatment the numerator is called the sum of square due to treatment SSTR divided by degrees of freedom.

(Refer Slide Time: 13:34)



Between-Treatments Estimate of Population Variance σ^2

- Mean Square due to Treatments (MSTR)

$$\text{MSTR} = \frac{\sum_{j=1}^{k} n_j(\bar{x}_j - \bar{\bar{x}})^2}{k-1}$$

Where:

k = number of groups

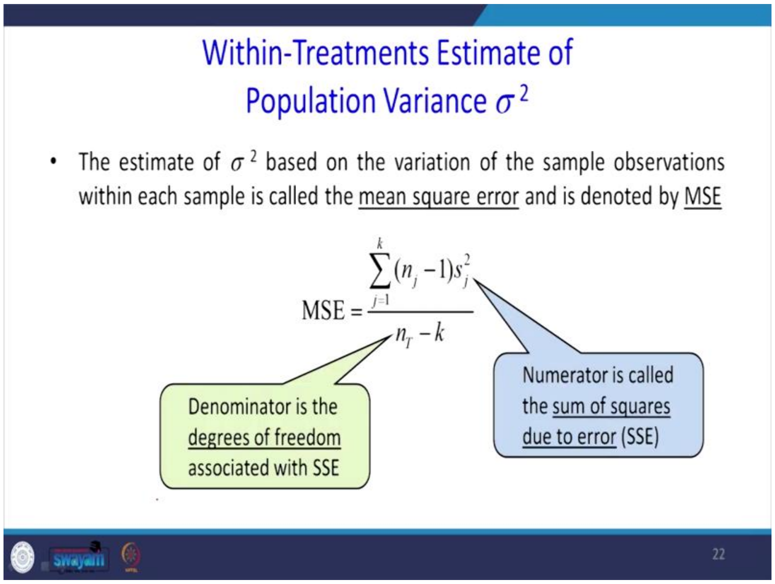
n_j = sample size from group j

$\bar{x}_j$ = sample mean from group j

$\bar{\bar{x}}$ = grand mean (mean of all data values)

Between treatment estimation of population variance Sigma square the mean square due to treatment that formula which you have seen previous slides so what is the meaning of this k, k is the number of groups, k number of columns nj is the sample size from Group j xj bar is the sample mean from Group j x double bar is the grand mean, mean of all data values over all mean.

(Refer Slide Time: 13:58)



Within-Treatments Estimate of Population Variance σ^2

- The estimate of σ^2 based on the variation of the sample observations within each sample is called the mean square error and is denoted by MSE

$$\text{MSE} = \frac{\sum_{j=1}^{k} (n_j - 1)s_j^2}{n_T - k}$$

Denominator is the degrees of freedom associated with SSE

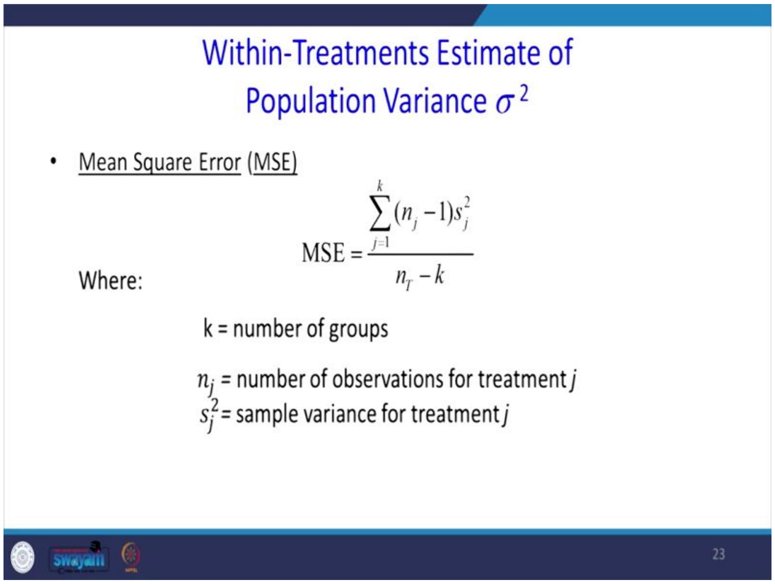
Numerator is called the sum of squares due to error (SSE)

Next we will see within treatment estimate our population variance. The estimator of Sigma square based on the variation of the sample observations within each sample this is more important term within each sample is called mean squared error is denoted by MSE. So, mean square error how we are doing that one nj in column j how many element is there minus one that is our degrees of freedom sj square.

Actually how it has come you see if we want to know sj square what is a formula Sigma of X minus for example X is a whole square divided by n - 1 so this instead of writing numerator that can be written as y sj square into n - 1 that is why it is written nj - 1 is j squared or nt is denominator is the degrees of freedom associated with error sum of square I will tell you that.

What is the n T in the next slide K is the number of groups n T is the number of treatment here this n T is nothing but the overall degrees of freedom. From the overall degrees of freedom if you subtract the degrees of freedom for between the columns then you will get either degrees of freedom for your SSE that is error sum of square. In our previous example we might have seen the n T is 9 – 1, 8 and the K is there was 3 columns, so it is 2 it was 6 degrees of freedom in our previous problem for MSE.

(Refer Slide Time: 15:43)



Within-Treatments Estimate of Population Variance σ^2

- Mean Square Error (MSE)

$$\text{MSE} = \frac{\sum_{j=1}^{k}(n_j - 1)s_j^2}{n_T - k}$$

Where:

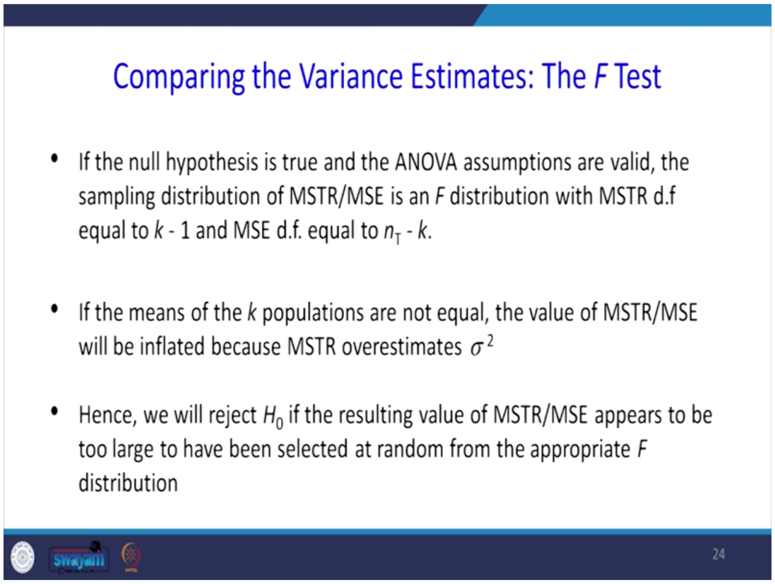
k = number of groups

n_j = number of observations for treatment *j*

s_j^2 = sample variance for treatment *j*

The formula for mean squared error is sigma of j equal to 1 to k n j - 1 is Z square by n T - 1 where n T is total number of observations where k is number of groups.

(Refer Slide Time: 16:00)



Comparing the Variance Estimates: The *F* Test

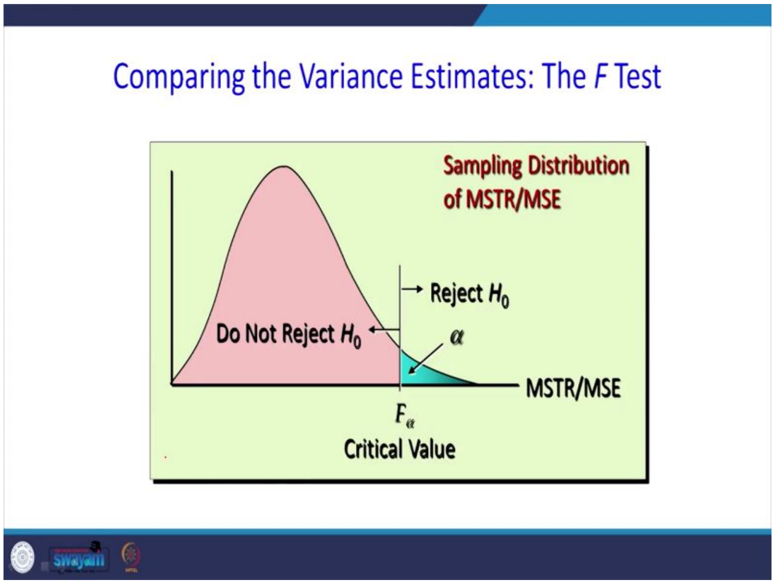
- If the null hypothesis is true and the ANOVA assumptions are valid, the sampling distribution of MSTR/MSE is an *F* distribution with MSTR d.f equal to k - 1 and MSE d.f. equal to $n_T - k$.
- If the means of the k populations are not equal, the value of MSTR/MSE will be inflated because MSTR overestimates σ^2
- Hence, we will reject H_0 if the resulting value of MSTR/MSE appears to be too large to have been selected at random from the appropriate *F* distribution

Comparing the variance estimates that is F test if the null hypotheses are true and ANOVA assumptions are valid the sampling distribution of MST are divided by MSE is an F distribution with MSTR degrees of freedom is equal to k - 1 that is number of column minus 1 and MSE a degrees of freedom is n T - 1 n T is total number of sample minus k is number of groups. If the means of the K populations are not equal the value of MSTR divided by MSE will be inflated because MSTR overestimate Sigma square.

So, what is the meaning of this one is we are finding you have the value of F is MSTR divided by MSE there are 2 possibility it may be equal 1 or less than 1 or greater than 1. If it is equal to 1 what is the meaning is variance due to treatment is equal to variance due to individual error. If it is greater than 1 the variance due to treatment is more when compared to within the error. When it become less than 1 if the MSE that is error due to individual differences more when compared to treatment then it will become less than 1.

So, you see this one if the mean of the k populations are not equal the value of MSTR by MSE will be inflated because the MSTR overestimate Sigma square. Hence we will reject because F l become very big when F l is very big we will reject H 0 if the resulting value of MSTR by MSE appears to be too large to have been selected at random from appropriate F distribution.

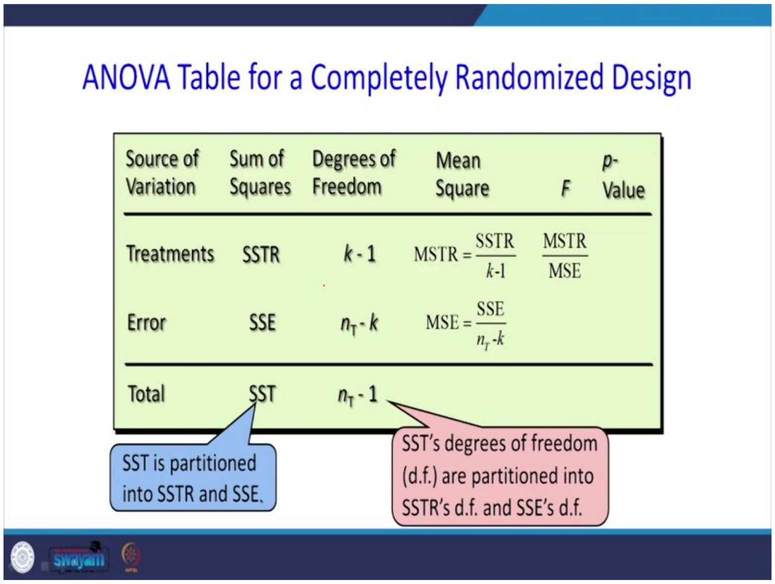
(Refer Slide Time: 17:52)



Comparing the Variance Estimates: The F Test

This is situation so what will happen when F is bigger number or obviously will be landing on the rejection site will reject null hypothesis. When you reject a null hypothesis we will say mu1 mu2 mu3 this was your null hypothesis, alternative hypothesis is mu 1 not equal to mu 2 not equal to me 3. So, when you reject null hypothesis we can conclude that this means are not equal and one more thing this is the F distribution it is not normal distribution it is a right skewed distribution.

(Refer Slide Time: 18:35)



ANOVA Table for a Completely Randomized Design

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F	p-Value
Treatments	SSTR	k - 1	$\text{MSTR} = \frac{\text{SSTR}}{k\text{-}1}$	$\frac{\text{MSTR}}{\text{MSE}}$	
Error	SSE	n_T - k	$\text{MSE} = \frac{\text{SSE}}{n_T\text{-}k}$		
Total	SST	n_T - 1			

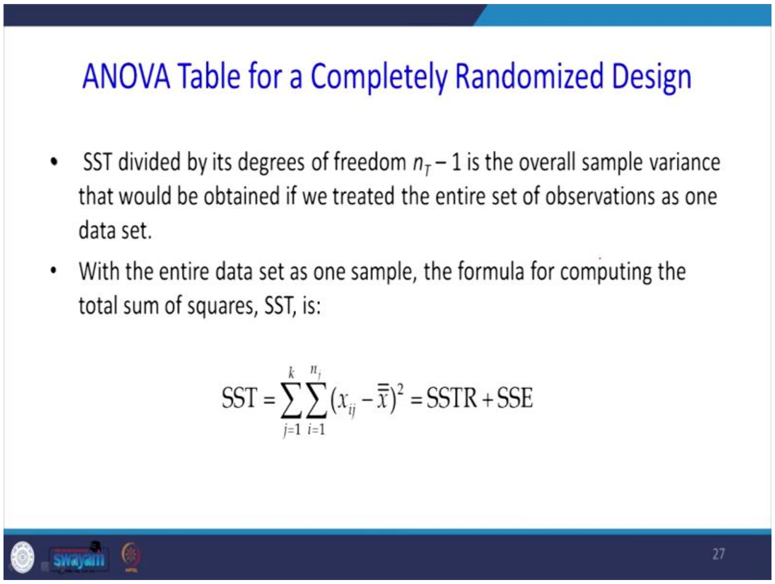
SST is partitioned into SSTR and SSE.

SST's degrees of freedom (d.f.) are partitioned into SSTR's d.f. and SSE's d.f.

This is the ANOVA table setup so what will be written sources of variation. So, there may be variation may be due to treatment variance due to error, so sum of square is sum of square treatment error sum of square. Here the deal is of freedom is K - 1 here n T - k generally if you assess the degrees of freedom is n T - 1 and T is total number of elements is 1 when you subtract this n 2 - 1 - k - 1 that will give you n T - K so MSTR is nothing but we have to divide this SSTR a little bit corresponding degrees of freedom. so, it will become mean treatment sum of square.

When you divide by SSE you total by corresponding degrees of freedom mean error sum square. So, the ratio of you see a MSTR divided by MSE always in the denominator there should be error term because when you go for 2-way ANOVA be able to remember that the denominator always there will be error term then we can find out corresponding p value this is what we have done previously when you are explaining my first example.

(Refer Slide Time: 19:41)



ANOVA Table for a Completely Randomized Design

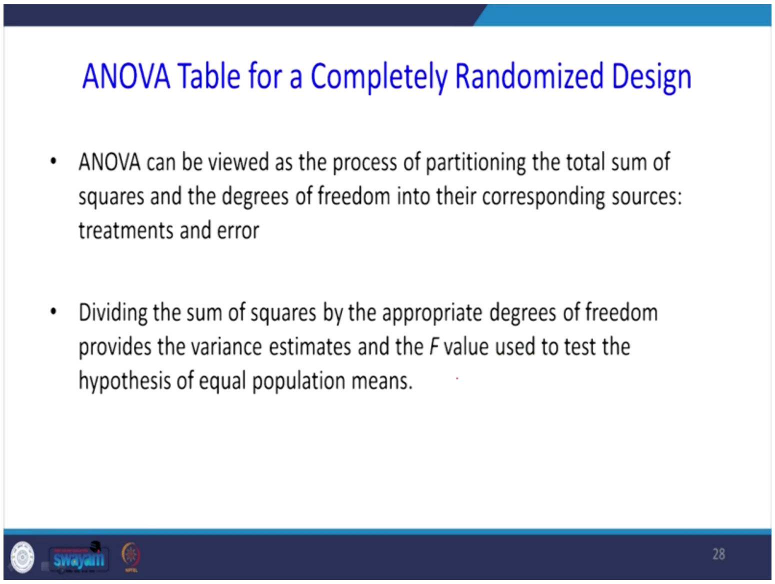
- SST divided by its degrees of freedom $n_T - 1$ is the overall sample variance that would be obtained if we treated the entire set of observations as one data set.
- With the entire data set as one sample, the formula for computing the total sum of squares, SST, is:

$$SST = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_{ij} - \bar{\bar{x}})^2 = SSTR + SSE$$

Generally what is happening this SST divided by its degrees of freedom nT - 1 is the overall sample variance that would be obtained if you treated the entire setup observation as one data set right when you divide this SST by corresponding degrees of freedom that is overall variance. With entire data set as a one sample the formula for computing the total sum of square is SST is sigma j equal to 1 to k sigma i equal to 1 to n say X ij - X double bar square.

So, this total sum of square can be splitted into 2 part one is treatment sum of square and error sum of square. If this treatment sum of square is dominating even without going further test we can say that there is a influence of treatment on the response variable.

(Refer Slide Time: 20:35)

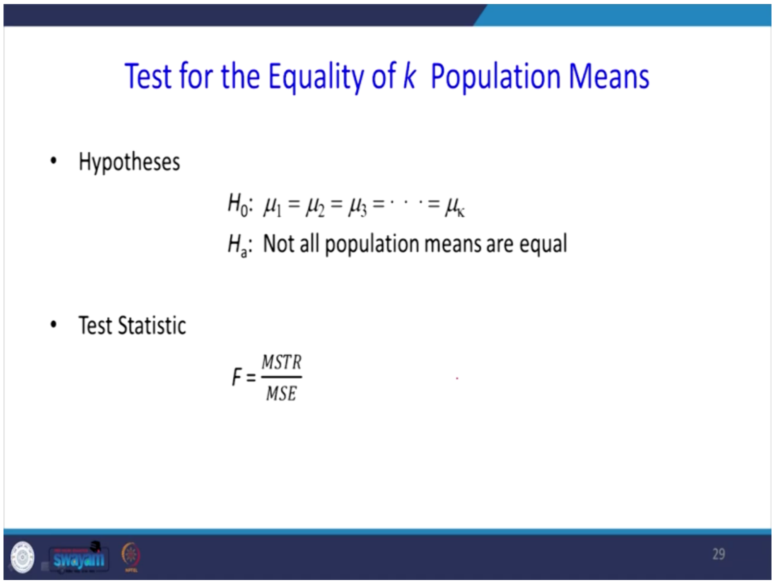


ANOVA Table for a Completely Randomized Design

- ANOVA can be viewed as the process of partitioning the total sum of squares and the degrees of freedom into their corresponding sources: treatments and error
- Dividing the sum of squares by the appropriate degrees of freedom provides the variance estimates and the *F* value used to test the hypothesis of equal population means.

ANOVA can be viewed as the process of partitioning the total sum of square and the degrees of freedom into their corresponding sources that is treatment and error. Dividing the sum of square by the appropriate degrees of freedom provides the variance estimates and the F value used to test the hypothesis of equal population means.

(Refer Slide Time: 21:02)



Test for the Equality of *k* Population Means

- Hypotheses

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

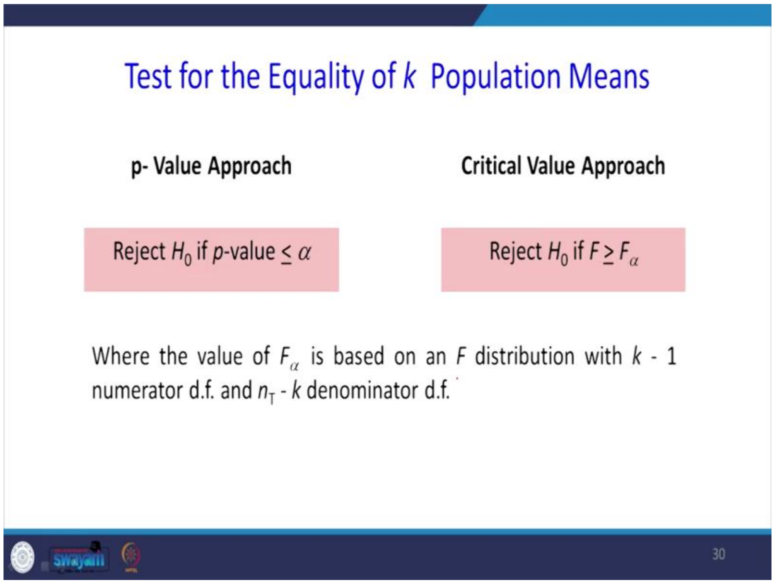
H_a: Not all population means are equal

- Test Statistic

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

What is the hypothesis the null hypothesis as usual mu1 equal to mu2 equal to mu 3, alternative hypothesis not all population means are equal the test statistic is the ratio of mean treatment sum of square root everybody mean error sum of square.

(Refer Slide Time: 21:19)



Test for the Equality of k Population Means

p- Value Approach

Reject H_0 if p-value $\leq \alpha$

Critical Value Approach

Reject H_0 if $F \geq F_\alpha$

Where the value of F_α is based on an F distribution with k - 1 numerator d.f. and n_T - k denominator d.f.

The p-value approach as usual for hypothesis testing also if the p-value is less than or equal to alpha we have to reject a null hypothesis. If you are using critical value the F value is greater than your value which you got from the table that also we have to reject our null hypothesis. In this class what we have seen we have taken one problem that is problem we have solved with help of Python then I have explained the theoretical background behind this ANOVA.

Then I have explained what is the total sum of square then what is the treatment sum of square than error sum of square. Then what is the degrees of corresponding degrees of freedom. In the next class will take extension of these classes once we reject a null hypothesis we have to say which 2 means are equal or not equal. So, that analysis is post hoc analysis we will continue the next lecture with the new topic of post hoc analysis in ANOVA, thank you very much.