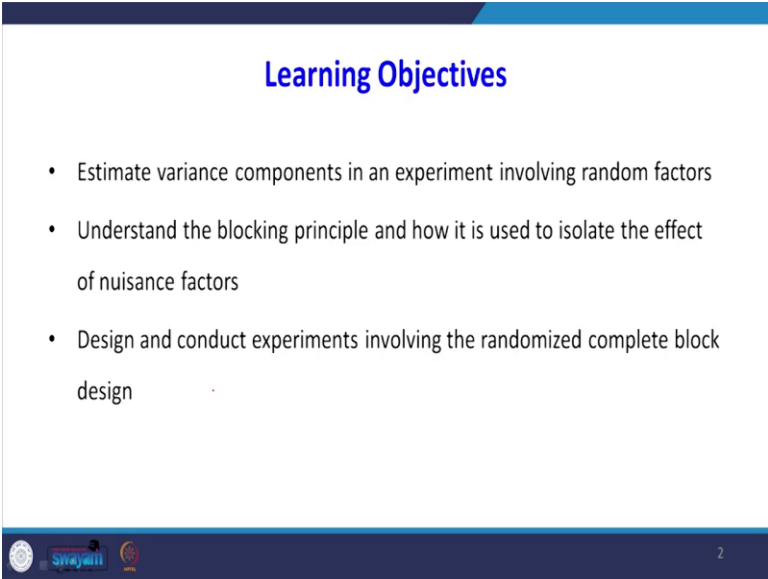


**Data Analytics with Python**  
**Prof. Ramesh Anbanandam**  
**Department of Management Studies**  
**Indian Institute of Technology – Roorkee**

**Lecture – 26**  
**Randomize Block Design (RBD)**

Dear students, the previous class we have seen one way Anova that is Completely Randomized Design, we call to CRD. In this class, we will see another technique called Randomized Block Design.

**(Refer Slide Time: 00:42)**



**Learning Objectives**

- Estimate variance components in an experiment involving random factors
- Understand the blocking principle and how it is used to isolate the effect of nuisance factors
- Design and conduct experiments involving the randomized complete block design

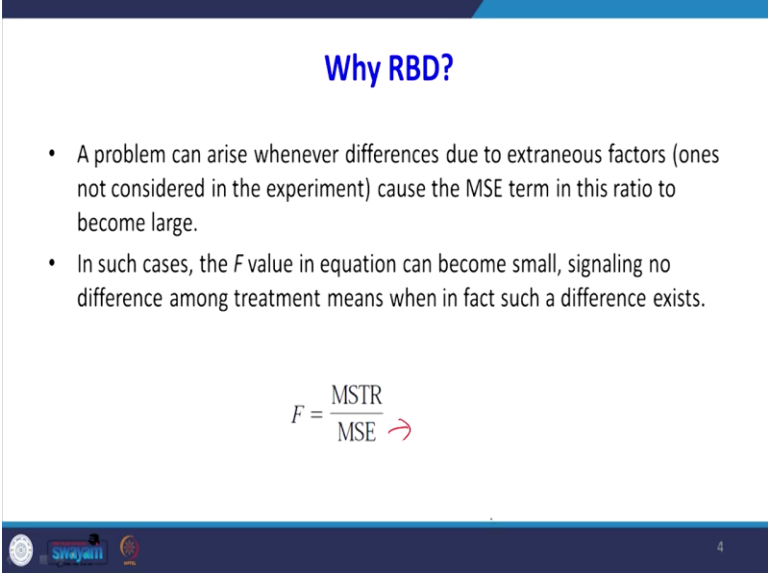
2

The class objectives are estimate the various components in experiment involving random factors, what will happen in Anova we are considering some factors. We are saying that the effect of the factor, what is the effect of the variance. But unknowingly, there is a possibility that some more variable may influence our response variable so that unknown variable and variance due to that unknown variables are going to remove it, then we are going to do the analysis.

Then, will see what is the effect of that one? Then, understand the blocking principle and how it is used to isolate the effect of nuisance factors. So, what you are doing here in Randomized block design. We are going to we are going to isolate the effect of nuisance factors then, design and conduct experiment involving Randomized Block design. A completely randomized design CRD

is useful when the experimental units are homogeneous. If the experiment units are heterogeneous blocking is often used to form homogeneous groups.

**(Refer Slide Time: 01:50)**



The slide is titled "Why RBD?" in blue text. It contains two bullet points: "A problem can arise whenever differences due to extraneous factors (ones not considered in the experiment) cause the MSE term in this ratio to become large." and "In such cases, the F value in equation can become small, signaling no difference among treatment means when in fact such a difference exists." Below the bullet points is the formula  $F = \frac{MSTR}{MSE}$  with a red arrow pointing to the MSE term. The slide has a dark blue header and footer with logos on the left and a small number '4' on the right.

**Why RBD?**

- A problem can arise whenever differences due to extraneous factors (ones not considered in the experiment) cause the MSE term in this ratio to become large.
- In such cases, the  $F$  value in equation can become small, signaling no difference among treatment means when in fact such a difference exists.

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

Why we have to go for RBD Randomized block design? A problem can arise whenever difference is due to extraneous factors that is, once not consider in the experiment cause the mean squared error term in this ratio to become large. What will happen? Due to that nuisance factor, the value of mean squared error will become very high. In such cases,  $f$  value in equation can become very small.

Signaling, no difference among treatment means when in fact such differences exist. So what will happen here in the MSE, there may be some error terms which are due to external factors. So we are going to find out how much error is due to external factor that we are going to remove it. Then, we are going to conduct the  $F$  Value.

**(Refer Slide Time: 02:45)**

## Randomized block design

- Experimental studies in business often involve experimental units that are highly heterogeneous; as a result, randomized block designs are often employed.
- Blocking in experimental design is similar to stratification in sampling.

Experimental studies in business often involve experimental units that are highly heterogeneous as a result Randomized block designs are often employed. Blocking in experimental design is similar to certificate stratification in sample. In stratification in sampling what we are doing? We are if the samples are heterogeneous based on certain criteria we are grouping, we are stratifying that sample, so that each strata will have homogeneous sample.

**(Refer Slide Time: 03:21)**

## Randomized block design


- Its purpose is to control some of the extraneous sources of variation by removing such variation from the MSE term.
- This design tends to provide a better estimate of the true error variance and leads to a more powerful hypothesis test in terms of the ability to detect differences among treatment means.

Here also, it is similar to stratification sampling. Its purpose is to control some of the external sources of variation by removing such variation from the MSE term. That is mean square error term. This design tends to provide a better estimate of the true error variance and leads to more powerful hypothesis test in terms of the ability to detect differences among treatment means.

(Refer Slide Time: 03:47)

### Air Traffic Controller Stress Test

- A study measuring the fatigue and stress of air traffic controllers resulted in proposals for modification and redesign of the controller's work station.
- After consideration of several designs for the work station, three specific alternatives are selected as having the best potential for reducing controller stress
- The key question is: To what extent do the three alternatives differ in terms of their effect on controller stress?



7

We will take one sample example. This sample example is Air traffic controller stress test. Why this Air Traffic Controller is he has to schedule various aircraft what time it has to be landed, what time it has to take off. So, he is the person who has to allocate different slots for, for landing and takeoff. So this job is very stressful job. We will see one problem on this one. A study measuring the stress of Air traffic controller resulted in a proposal for modification and redesign of controller's workstation.

So, what they are planning? They are going to redesign, the work station because this sometime the workstation may influence, may affect the stress level. If it is the workstation is very narrow people are get stressed more ok. After consideration of several designs for the workstation, 3 specific alternatives are selected, as having the best potential of reducing controllers stress. They have identified the three alternatives.

The key question is to what extent do the three alternatives differ in terms of their effect on controller stress? So we are going to see to what extent they are different, different workstation design is going to affect the stress of the Air traffic controller.

(Refer Slide Time: 05:16)

## Air Traffic Controller Stress Test

- In a completely randomized design, a random sample of controllers would be assigned to each work station alternative.
- However, controllers are believed to differ substantially in their ability to handle stressful situations.
- What is high stress to one controller might be only moderate or even low stress to another.
- Hence, when considering the within-group source of variation (MSE), we must realize that this variation includes both random error and error due to individual controller differences.
- In fact, managers expected controller variability to be a major contributor to the MSE term.

In a completely Randomized design a random sample of controllers would be assigned to each workstation alternative. Generally we will assign. However, Controllers are believed to differ substantially. It is not what we are assuming because the sample is not homogeneous because different controllers are affected by different level of workstation design. So, the controllers are believed to differ substantially in their ability to handle stressful situations.

What is high stress to one controller might be only moderate or even, low stress to another. So, what is happening? The sample is not homogeneous. Hence, when considering the within group sources of variation, that is MSE, let us call it means square error then, we must realize that this variation includes both random error and error due to individual control differences. In fact managers expected controller variability to be a major contribution to the MSE term.

**(Refer Slide Time: 06:26)**

## A randomized block design for the air traffic controller stress test

		Treatments		
		System A	System B	System C
Blocks	Controller 1	15	15	18
	Controller 2	14	14	14
	Controller 3	10	11	15
	Controller 4	13	12	17
	Controller 5	16	13	16
	Controller 6	13	13	13

This is a set up. So what happened? There are three workstations design. We call it system, A system, B system, C. See this is a controller 1. So, in controller 1, when we put into system 1, the stress level is measured in terms of 15. So when controller 1, when he was subjected to work design 2 workstation design B he was expecting the stress of the distance is measured in terms of a questionnaire, so the 15 is the score, higher the score, higher the stress.

So controller 1, controller 2, controller 3, controller 4 controller, there are 6 controller. Since each controllers are different, it there not homogeneous we are going to block it.

(Refer Slide Time: 07:08)

## Solving this example using ANOVA in python

```
ANOVA

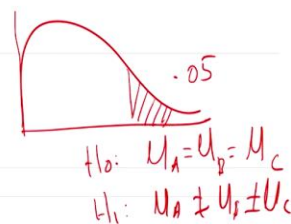
In [20]: data = pd.melt(df.reset_index(), id_vars=['index'], value_vars=['System A','System B','System C'])
data.columns = ['index', 'treatments', 'value']

In [21]: model = ols('value ~ C(treatments)', data=data).fit()
anova_table = sm.stats.anova_lm(model, typ=1)
anova_table

Out[21]:
```

	df	sum_sq	mean_sq	F	PR(>F)
C(treatments)	2.0	21.0	10.500000	3.214286	0.068903
Residual	15.0	49.0	3.266667	NaN	NaN

```
In [9]: # accept the null hypothesis
```



We solve these examples using Anova in Python. So, whether to import Pandas as pd, import numpy as np, import scipy, import statsmodels.api as sm, from statsmodels.formula.api, import ols. What I have done? The data which is in the table, which was in the previous slides, I have typed in an Excel. Then, I have imported that file name is RBD.xlsx so that I am going to save in the name of data frame. When I show the output this was system A, system B, system C.

So, I am using a melt command so data equal to `pd.melt(df.reset_index(), id_vars = ['index'], value_vars = ['system A', 'system B', 'system C'], data.columns = ['index', 'treatment', 'value'])`. This I also told you in the previous class. This melt command is used to bring all the values into column. One column is for treatment another column for values. So model equal to ols.

So, the value is the dependent variable tilde, See the treatments, data equal to that file name is data.fit. So anova\_table equal to `sm.stats.anova_lm`, lm is a linear model. (model, typ =1). Remember this Type 1 because when whenever there is a two way anova you have to use type 2. So, Anova underscore table. What is happening? This error sum of square is 3.2. So, what is happening this value is more than 0.05.

So, we are accepting null hypothesis. What is the meaning of accepting null hypothesis? Here, so if it is a 0.05, so it is the P value. 06 we accepted null hypothesis. When I accept null hypothesis, what is the null hypothesis here? The level of stress is equal for different 3 workstation design. Suppose,  $H_0$  equal to work station, This is stress, average stress level for workstation A, workstation design A, this is B, C. So  $H_1: \mu_A \neq \mu_B \neq \mu_C$ . So, at present what I am concluding I did not block it what time concluding? There is no connection between workstation design and the, and their average level of stress.

**(Refer Slide Time: 10:02)**

### Summary of stress data for the air traffic controller stress test

Treatments Blocks	System A	System B	System C	Block total	Block means
Controller 1	15	15	18	48	$\bar{x}_1 = 16$
Controller 2	14	14	14	42	$\bar{x}_2 = 14$
Controller 3	10	11	15	36	$\bar{x}_3 = 12$
Controller 4	13	12	17	42	$\bar{x}_4 = 14$
Controller 5	16	13	16	45	$\bar{x}_5 = 15$
Controller 6	13	13	13	39	$\bar{x}_6 = 13$
Column Total	81	78	93	252	$\bar{x} = 252/18 = 14$

Next, what I am going to do is I am going to do blocking going back. So there are 3.2 there is error. Actually 49, 49 divided by 15 so this is sum of square error is 49 mean sum of square is 3.2. So, this 3.26 due to blocking effect I am going to remove or subtract certain level of variance this. Then again, I am going to conduct, let us see what is happening. So this was the given data slide.

(Refer Slide Time: 10:33)

### Summary of stress data for the air traffic controller stress test

- Treatment means

$$\bar{x}_1 = 81/6 = 13.5$$

$$\bar{x}_2 = 78/6 = 13$$

$$\bar{x}_3 = 93/6 = 15.5$$

So the treatment mean is there are three treatments that I am calling this system A system B system C. So,  $\bar{x}_1$  is 13.5,  $\bar{x}_2$  is 13,  $\bar{x}_3$  is 15.5 ok.



(Refer Slide Time: 10:52)

**ANOVA TABLE FOR THE RANDOMIZED BLOCK DESIGN WITH  $k$  TREATMENTS AND  $b$  BLOCKS**

Sources of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F	P- value
Treatments	SS Treatments	$k-1$	$MS \text{ Treatments} = SSTR/k-1$	$MS \text{ Treatments} / MSE$	
Blocks	SS block	$(b-1)$	$MSBL = SSBL/b-1$		
Error	SSE	$(k-1)(b-1)$	$MSE = SSE/(k-1)(b-1)$		
Total	SST	$nb-1$			

We know this is our Anova setup. Look at the previous when there is the CRD in the completely Randomized design or one way Anova, there is no, there is no column blocking. Ok. There was only treatment and error. Now we are introducing the blocking so what is happening what are the degrees of freedom? Total number of element minus 1, degrees of freedom in treatment, there are  $k$  treatment  $k - 1$  blocking, there are  $b$  blocking  $b-1$ .

So how to find out the  $k-1$ ,  $b-1$  is  $n_T - 1 - (k-1) - (b-1)$ , so we get  $k-1$  and  $b-1$ . MSE treatment is SSE sum of square treatment divided by  $k - 1$ . MS mean square blocking equal to SSBL. Actually this data will not use that one. We will use only MS treatment by this MSE. SSE divided by  $k-1$  into  $b-1$ . Actually this much portions, we will remove this will be subtracted. Ok.

(Refer Slide Time: 12:03)

## RBD Problem

$x_{ij}$  = value of the observation corresponding to treatment  $j$  in block  $i$   
 $\bar{x}_{.j}$  = sample mean of the  $j$ th treatment  
 $\bar{x}_{i.}$  = sample mean for the  $i$ th block  
 $\bar{\bar{x}}$  = overall sample mean

So  $x_{ij}$  is the value of the observation corresponding to the treatment  $j$  in the block  $i$ .  $\bar{x}_{.j}$  is the sample mean of  $j$ th treatment.  $\bar{x}_{i.}$  is sample mean of  $i$ th block,  $\bar{\bar{x}}$  is overall sample mean.

(Refer Slide Time: 12:22)

## RBD Problem

**Step 1.** Compute the total sum of squares (SST).

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

**Step 1.**  $SST = (15 - 14)^2 + (15 - 14)^2 + (18 - 14)^2 + \dots + (13 - 14)^2 = 70$

**Step 2.** Compute the sum of squares due to treatments (SSTR).

$$SSTR = b \sum_{j=1}^k (\bar{x}_{.j} - \bar{\bar{x}})^2$$

**Step 2.**  $SSTR = 6[(13.5 - 14)^2 + (13.0 - 14)^2 + (15.5 - 14)^2] = 21$

What is the step 1? First, we will find out the SST that is a total sum of square. Total sum of squares summation  $i$  equal 1 to  $b$ , summation  $j$  equal 1 to  $j$   $x_{ij}$  - individual element - overall main whole square. Ok. So in that way, we are getting  $SST = 70$ . Then compute the sum of square due to treatment so there are 3 treatments so,  $b$  is number of replication treatment 1. That is 6 column 1 in 13.5 - overall mean 14 whole square + 6 is common be brought 6 everything is brought in this side. So,  $(13.0 - 14)^2 + (15.5 - 14)^2$  is 21.

(Refer Slide Time: 13:17)

**RBD Problem**

**Step 3.** Compute the sum of squares due to blocks (SSBL).

$$SSBL = k \sum_{i=1}^b (\bar{x}_{i.} - \bar{\bar{x}})^2$$

**Step 3.**  $SSBL = 3[(16 - 14)^2 + (14 - 14)^2 + (12 - 14)^2 + (14 - 14)^2 + (15 - 14)^2 + (13 - 14)^2] = 30$

**Step 4.** Compute the sum of squares due to error (SSE).

$$SSE = SST - SSTR - SSBL$$

**Step 4.**  $SSE = 70 - 21 - 30 = 19 //$

17

So, 3<sup>rd</sup> step is compute the sum of square due to blocks. Due to blocks is there is a k treatment ok, so  $\bar{x}_{i.}$  minus  $\bar{\bar{x}}$  double bar so that row wise what is the mean? Everywhere there are 3 treatment 3 this one. So, 16 -14 how we got the 16, going back to this 16,  $(16 - 14)^2 + (14 - 14)^2$  square and so on.  $(16 - 14)^2 + (14 - 14)^2 + (12 - 14)^2 + (14 - 14)^2 + (15 - 14)^2 + (13 - 14)^2$  is equal to 30.

This much variance is due to blocking, this much sum of square to compute the sum of square due to error term. We know that from SST you have to subtract treatment sum of square minus block in sum of square. That will give  $SSE = 70 - 21 - 30$  is 19. So this 19 is the true SSE because this SSB amount which is due to extraneous variable that we are noise variable. So, error due to this, we are removing this.

(Refer Slide Time: 14:47)

### ANOVA table for the air traffic controller stress test

Sources of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F	P- value
Treatments	21	2	10.5	10.5/1.9 =5.53	0.024
Blocks	30	5	6.0		
Error	19	10	1.9		
Total	70	17			

$$F_{.025} = 5.46 \text{ and } F_{.01} = 7.56.$$

Reject the null hypothesis

While finding SSE now what we are getting. Yeah, the, whatever value which are given I kept it here. So it is a 10.5 divided by 1.9. Even though we find this one will not used for calculation. So 5.59 to the value of P value is 0.024 if Alpha equal to 5 percentage, we have to reject the null hypothesis. Previously what has happened? When we do without blocking, we are accepted null hypothesis and going back see, we accepted null hypothesis without blocking.

After blocking, our decision has completely changed. So what happened? We have rejected the null hypothesis.

(Refer Slide Time: 15:34)

### Solving RBD example using python

```
In [1]: import pandas as pd
import numpy as np
import scipy
import statsmodels.api as sm
from statsmodels.formula.api import ols
```

```
In [4]: df = pd.read_excel('RBD.xlsx')
df
```

```
Out[4]:
```

	System A	System B	System C
0	15	15	18
1	14	14	14
2	10	11	15
3	13	12	17
4	16	13	16
5	13	13	13

We will do with the help of python import Pandas as pd, Import numpy as np, import scipy, import statsmodel.api as sm, from statsmodel.formula.api, import ols. ols is ordinary least square method because the regression and Anova is like two sides of the same coin. The sequence of learning Regression and Anova is first you have to learn Anova then you have to learn regression because there is a close relationship that I will see after this lecture is over after 2 lectures will go for Regression Analysis, I have imported.

**(Refer Slide Time: 16:13)**

### Solving RBD example using python


```
In [20]: data = pd.melt(df.reset_index(), id_vars=['index'], value_vars=['System A', 'System B', 'System C'])
data.columns = ['blocks', 'treatments', 'value']
```

```
In [22]: model = ols('value ~ C(block)+ C(treatments)', data=data).fit()
anova_table = sm.stats.anova_lm(model, typ=1)
anova_table
```

```
Out[22]:
```

	df	sum_sq	mean_sq	F	PR(>F)
C(block)	5.0	30.0	6.0	3.157895	0.057399
C(treatments)	2.0	21.0	10.5	5.526316	<u>0.024181</u>
Residual	10.0	19.0	1.9	NaN	NaN

```
In [23]: # reject the null hypothesis
```


20

Ok There are 3 columns that I am using the melt Command so that I will bring all the values in two columns. One is for here there are 3 columns that are going to do the blocking. Blocks treatments and values so model equal to ols( 'value tilde C (block) + C( treatment)' you see that now there is a blocking that I have included. Previously, there is no this term C x blocking so that I could data close bracket dot fit.

So, sm.stats.anova\_lm (model, is a typ = 1), anova\_table. I am getting you see this one here is this is 0.024 so it is less than 0.025 I am rejecting the null hypothesis.

**(Refer Slide Time: 17:07)**

## Conclusion

- Finally, note that the ANOVA table shown in Table provides an  $F$  value to test for treatment effects but *not* for blocks.
- The reason is that the experiment was designed to test a single factor—work station design.
- The blocking based on individual stress differences was conducted to remove such variation from the MSE term.
- However, the study was not designed to test specifically for individual differences in stress.

So what we are concluding. Finally note that Anova table shown in the table provides  $f$  value test for treatment effect but not for the blocks. The reason is that experiment was designed to test a single factor workstation design. The blocking based on the individual stress, stress differences was conducted to remove such variation from the MSE term. However, the study was not designed to test specifically for individual differences in stress.

What is happening here is, the blocking exactly what you are doing? The error due to blocking is removed while finding the influence of workstation design on stress level.

**(Refer Slide Time: 17:51)**

## Problem 2: RBD

- An experiment was performed to determine the effect of four different chemicals on the strength of a fabric.
- These chemicals are used as part of the permanent press finishing process.
- Five fabric samples were selected, and a randomized complete block design was run by testing each chemical type once in random order on each fabric sample.
- The data are shown in Table.
- We will test for differences in means using an ANOVA with  $\alpha = 0.01$ .

We will go for one more problem will use this Randomized block design. We will go for one more problem. An experiment was performed to determine the effect of four different Chemicals on strength of your fabric. These Chemicals are used as a part of permanent press finishing process. 5 fabric samples were selected. And a Randomized complete block design was run by testing each chemical type once in a random order on each fabric sample.

The data is shown in the table in the next slide. We will test the difference in using Anova with Alpha equal to 1 percentage.

(Refer Slide Time: 18:29)

**Problem 2: RBD**

- Table: Fabric Strength Data—Randomized Complete Block Design

Chemical Type	Fabric Sample					Treatment Totals	Treatment Averages
	1	2	3	4	5	$y_{i\cdot}$	$\bar{y}_{i\cdot}$
1	1.3	1.6	0.5	1.2	1.1	5.7	1.14
2	2.2	2.4	0.4	2.0	1.8	8.8	1.76
3	1.8	1.7	0.6	1.5	1.3	6.9	1.38
4	3.9	4.4	2.0	4.1	3.4	17.8	3.56
Block totals $y_{\cdot j}$	9.2	10.1	3.5	8.8	7.6	39.2( $y_{\cdot\cdot}$ )	
Block averages $\bar{y}_{\cdot j}$	2.30	2.53	0.88	2.20	1.90		1.96( $\bar{y}_{\cdot\cdot}$ )

This was the table. What says this? Different chemical type is there, different fabric samples are there. The replication is five because the same after adding chemical Type 1 when we conduct the fabric strength, we have conducted 5 samples. This was the row mean, this was the row average.

(Refer Slide Time: 18:49)

## Anova using jupyter

```
In [3]: df = pd.read_excel('rbd2.xlsx')
df
Out[3]:
```

	chem1	chem2	chem3	chem4
0	13	22	18	39
1	16	24	17	44
2	05	04	06	20
3	12	20	15	41
4	11	18	13	34

```
In [4]: data = pd.melt(df.reset_index(), id_vars='index', value_vars=['chem1', 'chem2', 'chem3', 'chem4'])
data.columns = ['index', 'treatments', 'value']

In [6]: model = ols('value ~ C(treatments)', data=data).fit()
aov_table = sm.stats.anova_lm(model, typ=1)
aov_table
Out[6]:
```

	df	sum_sq	mean_sq	F	PR(>F)
C(treatments)	3.0	18.044	6.014667	12.589569	0.000176
Residual	16.0	7.644	0.477750	NaN	NaN

What will you do? I have typed this data in excel, in excel file RBD2.xlsx. So this was the data so using melt coming and going to bring into the two variables. One is on value that is a response variable and other one is treatment. This you have to type as it is. That is the purpose of this pd.melt. So when I am running model = ols, a value is the dependent variable tilde, is the treatment is the independent variable.

Data is equal to data because this data is the way I have taken this after using melt command is the data. The file name is data so I am using data, data.fit and also another variable\_table = sm.stats.anova\_lm (model, is a typ = 1). So what is happening? You see that mean this is 0.4777. Here we did not do the blocking. We will do the blocking and what is happening to you? We are rejecting the null hypothesis because the probability is less than 0.01.

**(Refer Slide Time: 20:00)**



## Problem 2: RBD

- The sums of squares for the analysis of variance are computed as follows:

$$\begin{aligned}
 SS_T &= \sum_{i=1}^4 \sum_{j=1}^5 y_{ij}^2 - \frac{y_{..}^2}{ab} \\
 &= (1.3)^2 + (1.6)^2 + \dots + (3.4)^2 - \frac{(39.2)^2}{20} = 25.69 \\
 SS_{\text{Treatments}} &= \sum_{i=1}^4 \frac{y_{i.}^2}{b} - \frac{y_{..}^2}{ab} \\
 &= \frac{(5.7)^2 + (8.8)^2 + (6.9)^2 + (17.8)^2}{5} - \frac{(39.2)^2}{20} = 18.04
 \end{aligned}$$

So we are finding SST. SS treatment is see this formula is so comfortable for using calculator. So  $y_{ij}$  whole square minus  $y$  dot these notations, already I have explained. What is  $a$ ?  $a$  is the number of treatment  $b$  is the number of blocks? So this is SST is 25.69 is the total sum of square treatment sum of square is 18.04.

(Refer Slide Time: 20:27)

## Problem 2: RBD

$$\begin{aligned}
 SS_{\text{Blocks}} &= \sum_{j=1}^5 \frac{y_{.j}^2}{a} - \frac{y_{..}^2}{ab} \\
 &= \frac{(9.2)^2 + (10.1)^2 + (3.5)^2 + (8.8)^2 + (7.6)^2}{4} - \frac{(39.2)^2}{20} = 6.69 \\
 SS_E &= SS_T - SS_{\text{Blocks}} - SS_{\text{Treatments}} \\
 &= 25.69 - 6.69 - 18.04 = 0.96
 \end{aligned}$$

SS block is  $y_{.j}$  whole square minus  $y$  double dot whole square divided by  $ab$ . So, this 6.69, is the error term, you see that of finding SST. Total sum of square minus sum of square due to blocking that I am subtracting that is due to treatment so I am getting 0.96. So this is a true error without having blocking effect.

(Refer Slide Time: 20:54)

Problem 2: RBD					
• Analysis of Variance for the Randomized Complete Block Experiment					
Sources of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F	P- value
Chemical types (Treatments)	18.04	3	6.01	75.13	4.79 E-8
Fabric samples (Blocks)	6.69	4	1.67		
Error	0.96	12	0.08		
Total	25.69	19			

So what is happening? The mean square is here 0.08. So, you go back. What was the mean square without blocking? Yah, you see that it was without blocking it is 0.47 now it is .08. The mean square error is removed because we have removed the error due to blocking. So here the value of F also when you can compare it, it is significantly high. 75.13 that is a more chances for rejection. Previously, what is the F value. I am going back. Here the values is 12.12.58 now F value is 75.13. You are certainly you can say that you will reject your null hypothesis.

(Refer Slide Time: 21:50)

Conclusion	
<ul style="list-style-type: none"><li>The ANOVA is summarized in the previous table</li><li>Since <math>f_0 = 75.13 &gt; f_{0.01,3,12} = 5.95</math> (the <math>P</math>-value is <math>4.79 \times 10^{-8}</math>), we conclude that there is a significant difference in the chemical types so far as their effect on strength is concerned.</li></ul>	

Your Anova is summarised in the previous table. Since  $f$  equal to 75.13 which is greater than the table value that is a 5.95 which we got from the table, we have done Anova also. So that is the P

value is very low we conclude there is a significant difference in the chemical types so far as their effect of the strength is concerned.

(Refer Slide Time: 22:11)

### Python code for problem 2

```
In [2]: import pandas as pd
import statsmodels.api as sm
from statsmodels.formula.api import ols
from statsmodels.stats.anova import anova_lm
```

```
In [3]: df = pd.read_excel('RBD2.xlsx')
```

```
In [4]: df
```

```
Out[4]:
```

	chem1	chem2	chem3	chem4
0	1.3	2.2	1.8	3.9
1	1.6	2.4	1.7	4.4
2	0.5	0.4	0.6	2.0
3	1.2	2.0	1.5	4.1
4	1.1	1.8	1.3	3.4

29

Previously, we have done a traditional way. Now we will use Python for doing the blocking that is doing the Randomized block design. import pandas as pd, import statsmodels.api as sm, from statsmodels.formula.api import ols from statmodels.stats.anova import anova\_lm. So, you save the file in the name df equal to pd.read\_excel (). df This was the output.

(Refer Slide Time: 22:41)

### Python code for problem 2

```
In [7]: data = pd.melt(df.reset_index(), id_vars=['index'], value_vars=['chem1','chem2','chem3','chem4'])
data.columns = ['Fabric samples', 'Chemical types', 'value']
data
```

```
Out[7]:
```

	Fabric samples	Chemical types	value
0	0	chem1	1.3
1	1	chem1	1.6
2	2	chem1	0.5
3	3	chem1	1.2
4	4	chem1	1.1
5	0	chem2	2.2
6	1	chem2	2.4
7	2	chem2	0.4
8	3	chem2	2.0
9	4	chem2	1.8
10	0	chem3	1.8
11	1	chem3	1.7
12	2	chem3	0.6
13	3	chem3	1.5
14	4	chem3	1.3
15	0	chem4	3.9
16	1	chem4	4.4
17	2	chem4	2.0
18	3	chem4	4.1

30

Again, you see that we are using melt command after giving the melt command the data has become this format. So what is happening? Fabric samples 01234, 01234 see, these are 1 group.

This is another group. This is another group. This is another group, another group. So, this is chemical 1, chemical 2, chemical3, chemical4 the purpose of this pd.melt command is for this purpose.

Now there are three columns. One is the fabric sample. So the value, value is dependent variable chemical type treatment is independent variable. Fabric sample, that is, blocking variables.

**(Refer Slide Time: 23:28)**

**Python code for problem 2**

```
In [11]: model = ols('value ~ C(Fabric) + C(Chemical)', data=data).fit()
         anova_table = sm.stats.anova_lm(model, typ=1)
         anova_table
```

Out[11]:

	df	sum_sq	mean_sq	F	PR(>F)
C(Fabric)	4.0	6.693	1.673250	21.113565	2.318913e-05
C(Chemical)	3.0	18.044	6.014667	75.894849	4.518310e-08
Residual	12.0	0.951	0.079250	NaN	NaN

31

Now past this model is equal to ols ('value tilde C (fabric), fabric is, is a blocking effect plus the chemical that is the treatment effect, data equal to data.fit. When you run this we are getting see the f value is which we got traditionally manual method. We got this one so we see that P value. So what we have done, we taken one problem, we have solved without blocking what was the status. In this problem we are rejecting then we go for blocking.

After blocking also we are rejecting. But the when you look at the value of f that is significantly, it has increased. So, what will happen without blocking you may conclude on things? You may accept null hypothesis, because the error term is very bigger. After blocking the error term become very less than you may reverse the decision. We can reject the null hypothesis. That is application of this blocking.

Dear Students, in this lecture, what we have seen just I am summarizing. We have seen what is randomized block design? We have seen what is the need when will we go for a Randomized block design. Then you have taken your problem that problem was solved without blocking and seen what was the result then the same problem with blocking. Then you have seen how the result has changed. Even without blocking also we are used python code we saw what is the result?

Then, with blocking also we have used Python code. Then we have seen what was the result? In this, what we have done? We have taken two problems for both the problems we solved with blocking and without blocking. In the next class, we are going to another type of Anova that is a two way Anova. Thank you very much.