



IIT ROORKEE



NPTEL ONLINE
CERTIFICATION COURSE

this is

this is

RBD

Dr. A. Ramesh
DEPARTMENT OF MANAGEMENT
IIT ROORKEE



One way ANOVA

CRD

~~CRD~~

Learning Objectives

ANOVA - Sum of squares
effect variance

other var may effect

- 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 } noise
- Design and conduct experiments involving the randomized complete block design

all Remove effect of other var. CRD used when experiment units homogenous

If exp design hetero given \rightarrow RBD.

Randomized Block Design

Blocking ↓
form homogeneous
groups

- A completely randomized design (CRD) is useful when the experimental units are homogeneous
- If the experimental units are heterogeneous, **blocking** is often used to form homogeneous groups

RBD \Rightarrow turn homogeneous \rightarrow form homogeneous groups
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.

due to high MSE \rightarrow F will decrease \rightarrow overall effect on the decreases dep var / variance decreases

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

\hookrightarrow small \rightarrow large \hookrightarrow loss due to external factors
 \hookrightarrow indicate that no diff in means but in reality diff exist.

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.

this is

if sample heterogeneous
we group such that
each strata → homogeneous
sample

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.
- diff among treatment means*

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?



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.

Not
human
errors

Error

A randomized block design for the air traffic controller stress test

diff work station

Treatments

high speed

low speed

	System A	System B	System C
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

Blocks

Solving this example using ANOVA in python

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

} statsmodels

```
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

Sum Sum/3
48 16
42 14
36 12
42 14
45 15
39 13

Solving this example using ANOVA in python

ANOVA

```
[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
```

	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
```

Variable
if exists
Not due to
cause station
as per Ho

10 $k=3$ 3 groups

} mult

variable

model

amount

lm = linear model

Accept Ho

type 1 = one factor

Level of sig = equal for all cause stations
 $\mu_a = \mu_b = \mu_c$



Summary of stress data for the air traffic controller stress test

Treatments → Blocks ↓	System A	System B	System C	<u>Block total</u>	Block means
Controller 1	15	15	18	48	$\overline{x_{1.}} = 16$
Controller 2	14	14	14	42	$\overline{x_{2.}} = 14$
Controller 3	10	11	15	36	$\overline{x_{3.}} = 12$
Controller 4	13	12	17	42	$\overline{x_{4.}} = 14$
Controller 5	16	13	16	45	$\overline{x_{5.}} = 15$
Controller 6	13	13	13	39	$\overline{x_{6.}} = 13$
Column Total	81	78	93	252	$\overline{\overline{\overline{x}}} = 252/18 = 14$

Summary of stress data for the air traffic controller stress test

- Treatment means

System A $\rightarrow \bar{x}_1 = 81/6 = 13.5$
B $\rightarrow \bar{x}_2 = 78/6 = 13$
C $\rightarrow \bar{x}_3 = 93/6 = 15.5$



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

$$n_T - 1 = (k - 1) - (b - 1) \\ = n - k - b + 1$$

Sources of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F	P- value
Treatments	SS Treatments	$k - 1$ <i>same</i>	MS Treatments = $SSTR / k - 1$	MS Treatments / MSE	
<i>RRD</i> Blocks	<u>SS block</u>	$(b - 1)$ ✓	MSBL = $SSBL / b - 1$ <i>NOT USED</i>		
Error	SSE	$(k - 1)(b - 1)$	MSE = $SSE / (k - 1)(b - 1)$		
Total	SST	$n_T - 1$ <i>same</i>			

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 .

RBD Problem

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

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

✓ same.

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).

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

✓ same except block.

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

RBD Problem

$k \text{ treatment} = 8$

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

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

variance
due to

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

$\Rightarrow SS(BL)$ due to blocking

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

$$SSE = SST - SSTR - SSBL$$

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

\rightarrow due to noise.

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	<u>0.024</u>
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

$P < 0.05$ Reject H_0

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

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
```

formula different
C(block)

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	0.024181
Residual	10.0	19.0	1.9	NaN	NaN

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

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.

what should handle this situation

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 $\alpha = 0.01$.

→ 99

Problem 2: RBD

- Table: Fabric Strength Data—Randomized Complete Block Design

Chemical Type	Fabric Sample					Treatment Totals	Treatment Averages
	1	2	3	4	5	$y_{i.}$	$\bar{y}_{i.}$
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 $y_{.j}$	9.2	10.1	3.5	8.8	7.6	39.2($y_{..}$)	
Block averages $\bar{y}_{.j}$	2.30	2.53	0.88	2.20	1.90	1.96($\bar{y}_{..}$)	

Row avg

Anova using jupyter

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

Out[3]:

	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

understand the
experiment and choose
whether it needs RP or
RBD

```
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
```

] melt
] RBD No blocks.

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

Problem 2: RBD

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

$$\begin{aligned}SS_T &= \sum_{i=1}^4 \sum_{j=1}^5 y_{ij}^2 - \frac{y_{..}^2}{ab} \\&= (1.3)^2 + (1.6)^2 + \cdots + (3.4)^2 - \frac{(39.2)^2}{20} = 25.69\end{aligned}$$

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

Problem 2: RBD

$$SS_{\text{Blocks}} = \sum_{j=1}^5 \frac{y_{.j}^2}{a} - \frac{y_{..}^2}{ab}$$
$$= \frac{(9.2)^2 + (10.1)^2 + (3.5)^2 + (8.8)^2 + (7.6)^2}{4} - \frac{(39.2)^2}{20} = 6.69$$

$$SS_E = SS_T - SS_{\text{Blocks}} - SS_{\text{Treatments}}$$
$$= 25.69 - 6.69 - 18.04 = 0.96$$

*This will increase f
and \therefore Reduced Pval*

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	→ Enter it as 0.07	
Total	25.69	19			

F over 12.08

→ Enter it as 0.07

Conclusion

F_{α}

- The ANOVA is summarized in the previous table
- Since $f_0 = 75.13 > f_{0.01,3,12} = 5.95$ (the P -value is 4.79×10^{-8}), we conclude that there is a significant difference in the chemical types so far as their effect on strength is concerned.

Can be

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
```

Regular

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

import data

```
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

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
19	4	chem4	2.4

data preparation

id-
var

dep. var

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
```

RBD

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	75.894848	4.518310e-08
Residual	12.0	0.951	0.079250	NaN	NaN