STAT22200 Chapter 14

Yibi Huang

Chapter 14 Incomplete Block Designs

14.1 Balanced Incomplete Block Designs (BIBD)

Incomplete Block Designs

— A Brief Introduction to a Class of Most Useful Designs in Practice

Recall that for a randomized complete block design (RCBD) of g treatments, the size k of each block has to be g (or multiples of g). Under complete block design, each treatment occurs once and only once in each block, so that we can randomize and run all treatment combinations in each block.

However, in practice, it often happens that the size of available blocks is smaller than the numbers of treatments ($k < g$). We cannot apply every treatment to every block.

We then have **Incomplete Block Design (IBD)**. IBD makes analysis harder, but sometimes cannot be avoided.

The next best thing to a randomized complete block design (RCBD) is a **Balanced Incomplete Block Design (BIBD)**.

An Example of Unavoidable Incomplete Blocks

Eye irritation can be reduced with eyedrops. Three brands of eyedrops are to be compared for their ability to reduce eye irritation.

As there is a strong individual effect, subjects should be used as blocks.

If each subject can only be used during one treatment period, then the researchers must use one brand of drop in the left eye and another brand in the right eye. The design is restricted into blocks of size two.

The study is force into incomplete blocks, with

$$\begin{array}{ccc} k = 2 & < & 3 = g \\ \text{(block size)} & < & \text{(number of treatments)} \end{array}$$

Example — A Marketing Psychology Experiment

- Goal: comparing 5 commercial ads: A, B, C, D, E
- Response: subjects' rating of a commercial ad after watching it
- A subject can watch multiple ads. A subject is a block
- Can use a RCBD of all subjects can watch all the 5 ads
- However, subjects may lose patience after watching too many ads, and they may forget the first few ads they see. Their response will be less accurate.
- To ensure the quality of the response of subjects, we may restrict the number of ads each subject watch to, say, 3. The block size is limited to $k = 3$.

Subject

1	2	3	4	5	6	7	8	9	10
A	B	E	A	C	D	B	E	D	C
B	D	A	C	A	E	C	B	E	D
C	A	B	D	E	A	D	C	B	E

Balanced Incomplete Block Designs (BIBD)

BIBD is not balanced in the general sense that all treatment-block combinations occur equally often. Rather they are balanced in the looser sense by the criteria described below.

A **balanced incomplete block design** with

g treatments,

b blocks,

k as the size of each block,

r replications of each treatment,

is a design satisfying the following:

Incomplete:
- $k < g$.

Balanced:
- Each treatment appears at most once per block and has the same number of replicates r
- Each pair of treatments appear in a block the same number of times λ

Subject (Block)

1	2	3	4	5	6	7	8	9	10
A	B	E	A	C	D	B	E	D	C
B	D	A	C	A	E	C	B	E	D
C	A	B	D	E	A	D	C	B	E

The design of the marketing psychology study is a BIBD with

$$\begin{aligned} g &= \text{number of treatments} = 5 \\ b &= \text{number of blocks} = 10 \\ k &= \text{size of each block} = 3 \\ r &= \text{number replicates per treatment} = 6 \end{aligned}$$

The table below shows the blocks each treatment appears, verifying that each treatment appear $r = 6$ times.

	Block									
Treatment	1	2	3	4	5	6	7	8	9	10
A	√	√	√	√	√	√				
B	√	√	√				√	√	√	
C	√			√	√		√	√		√
D		√		√		√	√		√	√
E			√		√	√		√	√	√

BIBD requires each pair of treatments appears in a block the same number (λ) of times.

Subject (Block)									
1	2	3	4	5	6	7	8	9	10
A	B	E	A	C	D	B	E	D	C
B	D	A	C	A	E	C	B	E	D
C	A	B	D	E	A	D	C	B	E

The table below verifies that, each treatment pair appears $\lambda = 3$ times for the design above.

Treatment-pair	Block									
	1	2	3	4	5	6	7	8	9	10
AB	√	√	√							
AC	√			√	√					
AD		√		√		√				
AE			√		√	√				
BC	√						√	√		
BD		√					√		√	
BE			√					√	√	
CD				√			√			√
CE					√			√		√
DE						√			√	√

First Balancing Condition of BIBD

The five numbers that describe a BIBD: g, b, k, r, and λ are not arbitrary.

There might not exist an allocation b blocks of k units to g treatments that is a BIBD.

- If there are b blocks of size k each,
  then the total number of experimental units is $N = bk$.
- If there are g treatments, and each appears r times,
  then the total number of experimental units is $N = rg$.

Therefore in a BIBD, we must have

$$\boxed{N = bk = rg}.$$

Second Balancing Condition of BIBD

In a BIBD, every pair of treatments must appears in a block the same number of times, say λ times.

Observe that the total number of pairings involving treatment A equals

- $\lambda(g-1)$, since A may pair with any of the other $g-1$ treatments, and each pair appears in λ blocks.
- $r(k-1)$, since treatment A appears in r blocks. Within each of those blocks, there are $k-1$ pairs including A as the block size

The **second balancing condition**

$$\boxed{r(k-1) = \lambda(g-1)}$$

Given g treatments and b blocks of size k, one can show that a BIBD that with r replicates per treatment and each pair of treatments show in a block λ times exists if and only if

$$bk = rg \text{ and } r(k-1) = \lambda(g-1).$$

Example (Eyedrop): $g = 3$, $k = 2$.

- Is it possible to find a BIBD w/ $b = 5$ subjects (blocks)?
  No, as $r = bk/g = 2 \cdot 5/3 = 10/3$ is NOT an integer.
- Is it possible to find a BIBD w/ $b = 6$ subjects (blocks)?
  Yes, as $r = bk/g = 2 \cdot 6/3 = 4$ and $\lambda = \frac{r(k-1)}{(g-1)} = \frac{4(2-1)}{3-1} = 2$ are both integers

Example (Marketing Psychology): $g = 5$, $k = 3$.

- Is it possible to find a BIBD w/ $b = 5$ subjects (blocks)?
  No. $r = bk/g = 3 \cdot 5/5 = 3$ is an integer, but $\lambda = \frac{r(k-1)}{(g-1)} = \frac{3(3-1)}{5-1} = 6/4$ is NOT an integer.
- Is it possible to find a BIBD w/ $b = 10$ subjects (blocks)?
  Yes, as $r = bk/g = 10 \cdot 3/5 = 6$ and $\lambda = \frac{r(k-1)}{(g-1)} = \frac{6(3-1)}{5-1} = 3$ are both integers

Just like Latin Squares, it's not trivial to find a BIBD by oneself.

Appendix C.2 on p.609-615 of Oehlert's textbook gives a list of some BIBD plans for $g \leq 9$.

- **A BIBD can be replicated to conduct a larger study.** E.g., in the marketing psychology experiment, if we have $b = 20$ subjects (blocks) instead of 10, then we can do 2 repetitions of the BIBD below with $g = 5$, $k = 3$, $b = 10$, $r = 6$, $\lambda = 3$:

A	B	E	A	C	D	B	E	D	C
B	D	A	C	A	E	C	B	E	D
C	A	B	D	E	A	D	C	B	E

- **How to Do Randomization in BIBD?** One obvious randomization is to randomize subjects to columns, then randomize the order of treatments in each block based on the above design.

Models for BIBD

The model for BIBD looks familiar:

$$y_{ij} = \mu + \underset{\text{(treatment)}}{\alpha_i} + \underset{\text{(block)}}{\beta_j} + \underset{\text{(i.i.d. } N(0,\sigma^2))}{\varepsilon_{ij}}$$

for $i = 1, \ldots, g$, and $j = 1, \ldots, b$ with

$$\sum_{i=1}^{g} \alpha_i = \sum_{j=1}^{b} \beta_j = 0.$$

- additive model (no treatment-block interaction)
- Not all y_{ij} exist because of incompleteness
- Due to the incompleteness, just like *unbalanced* factorial designs. Thus, Type I Sum of squares will change with the order of terms in the model.

Parameter Estimates for BIBD

Let

$$I_{ij} = \begin{cases} 1, & \text{if treatment } i \text{ appears in block } j, \\ 0, & \text{otherwise.} \end{cases}$$

and define

$$Q_i = y_{i\bullet} - \frac{1}{k}\sum_j I_{ij} y_{\bullet j}, \qquad Q'_j = y_{\bullet j} - \frac{1}{r}\sum_i I_{ij} y_{i\bullet}$$

the least square estimates for μ, α_i, β_j are

$$\widehat{\mu} = \frac{y_{\bullet\bullet}}{N}, \qquad \widehat{\alpha}_i = \frac{kQ_i}{\lambda g}, \qquad \widehat{\beta}_j = \frac{kQ'_j}{\lambda b}$$

Remark: Can verify that $\sum_i Q_i = 0 \quad \Rightarrow \sum_i \widehat{\alpha}_i = 0$.

You won't be asked to estimate parameters manually for a BIBD.

ANOVA for BIBD (Type I Sum of Squares!)

Source	d.f.	SS	MS	F-value
Block	$b-1$	SS_{block}	MS_{block}	$(\text{MS}_{block}/\text{MSE})$
Treatment	$g-1$	SS_{trt}	MS_{trt}	$\text{MS}_{trt}/\text{MSE}$
Error	$N-g-b+1$	SSE	MSE	
Total	$N-1$	SS_{total}		

Let $$I_{ij} = \begin{cases} 1, & \text{if treatment } i \text{ appears in block } j, \\ 0, & \text{otherwise.} \end{cases}$$

Then $$SS_{total} = \sum_{i=1}^{g} \sum_{j=1}^{b} I_{ij}(y_{ij} - \overline{y}_{\bullet\bullet})^2$$

$$SS_{block} = k \sum_{j=1}^{b} (\overline{y}_{\bullet j} - \overline{y}_{\bullet\bullet})^2 \qquad \text{(unadjusted, Type I)}$$

$$SS_{trt} = \frac{k}{\lambda g} \sum_i Q_i^2 = \frac{\lambda g}{k} \sum_i \widehat{\alpha}_i^2 \quad \text{(adjusted for block, Type I \& II)}$$

$$SSE = SS_{total} - SS_{block} - SS_{trt}$$

For incomplete block designs, **always place Block ahead of Treatment in the ANOVA table. The SS$_{trt}$ will then be adjusted for Block and hence is Type II.**

Pairwise Comparisons

Estimate of $\alpha_{i_1} - \alpha_{i_2}$ is

$$\widehat{\alpha}_{i_1} - \widehat{\alpha}_{i_2} = \frac{k}{\lambda g}(Q_{i_1} - Q_{i_2})$$

- $\text{SE}(\widehat{\alpha}_{i_1} - \widehat{\alpha}_{i_2}) = \sqrt{\text{MSE}\left(\frac{2k}{\lambda g}\right)}$
- $t\text{-statistic} = \frac{\widehat{\alpha}_{i_1} - \widehat{\alpha}_{i_2}}{\text{SE}}$ with df = df of MSE

Example of BIBD — Problem 14.3 on. p.381-382

The State Board of Education has adopted basic skills tests for high school graduation. One of these is a writing test. The student writing samples are graded by professional graders, and the board is taking some care to be sure that the graders are grading to the same standard. We examine grader differences with the following experiment. There are 25 graders available. We select 30 writing samples at random, each writing sample will be graded by 5 graders. Thus each grader will grade $30 \times 5/25 = 6$ samples.

Data file: http://users.stat.umn.edu/~gary/book/fcdae.data/pr14.3

Questions of Interest:

- Did the 25 graders grade consistently with each other?
- How to adjust the scores if graders didn't grade consistently?
- If graders didn't grade consistently, can we identify the graders that were inconsistent with others?

Exam	Grader					Score					Exam	Grader					Score				
1	1	2	3	4	5	60	59	51	64	53	16	1	9	12	20	23	61	67	69	68	65
2	6	7	8	9	10	64	69	63	63	71	17	2	10	13	16	24	78	75	76	75	72
3	11	12	13	14	15	84	85	86	85	83	18	3	6	14	17	25	67	72	72	75	76
4	16	17	18	19	20	72	76	77	74	77	19	4	7	15	18	21	84	81	76	79	77
5	21	22	23	24	25	65	73	70	71	70	20	5	8	11	19	22	81	84	85	84	81
6	1	6	11	16	21	52	54	62	54	55	21	1	8	15	17	24	70	65	61	66	66
7	2	7	12	17	22	56	51	52	57	51	22	2	9	11	18	25	84	82	86	85	86
8	3	8	13	18	23	55	60	59	60	61	23	3	10	12	19	21	72	85	77	82	79
9	4	9	14	19	24	88	76	77	77	74	24	4	6	13	20	22	85	75	78	82	83
10	5	10	15	20	25	65	68	72	74	77	25	5	7	14	16	23	58	64	58	57	58
11	1	10	14	18	22	79	77	77	77	79	26	1	7	13	19	25	66	71	73	70	70
12	2	6	15	19	23	70	66	63	62	66	27	2	8	14	20	21	73	67	63	70	66
13	3	7	11	20	24	48	49	51	48	50	28	3	9	15	16	22	58	70	69	61	71
14	4	8	12	16	25	75	64	75	68	65	29	4	10	11	17	23	95	84	88	88	87
15	5	9	13	17	21	79	77	81	79	83	30	5	6	12	18	24	47	47	51	49	56

Here a exam is a writing sample.

- Which factor is the treatment factor? Graders or Exams?
  Graders.
- Which factor is the block factor? Graders or writing samples?
  Exams.
- Is this a BIBD?
  Yes, $g = 25$, $b = 30$, $k = 5$, $r = \frac{bk}{g} = 6$, $\lambda = \frac{r(k-1)}{g-1} = 1$.

$$\underset{\text{(score)}}{y_{ij}} = \mu + \underset{\text{(grader)}}{\alpha_i} + \underset{\text{(exam)}}{\beta_j} + \varepsilon_{ij}$$

As writing samples differ in levels, we expect β_j not all equal.

If graders were consistent, they should give the same score to the same writing sample, i.e., $\alpha_1 = \alpha_2 = \cdots = \alpha_{25}$

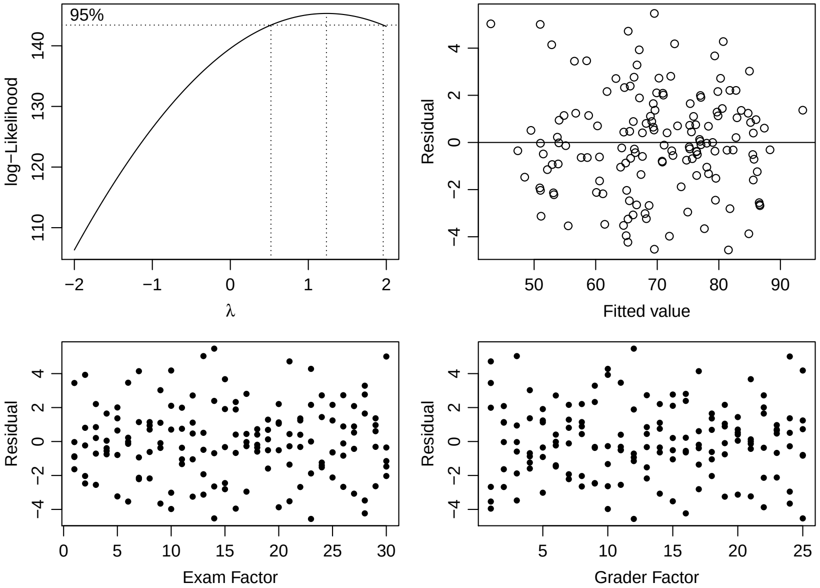
```
pr14.3 = read.table(
  "http://users.stat.umn.edu/~gary/book/fcdae.data/pr14.3", h=T)
pr14.3$EXAM = as.factor(pr14.3$exam)
pr14.3$GRADER = as.factor(pr14.3$grader)
```

```
> anova(lm(score ~ GRADER + EXAM, data=pr14.3))
          Df  Sum Sq Mean Sq F value    Pr(>F)    
GRADER    24  4073.1  169.71  23.659 < 2.2e-16 ***
EXAM      29 13342.0  460.07  64.138 < 2.2e-16 ***
Residuals 96   688.6    7.17                      

> anova(lm(score ~ EXAM + GRADER, data=pr14.3))
          Df  Sum Sq Mean Sq F value    Pr(>F)    
EXAM      29 16609.0  572.72 79.8424 < 2.2e-16 ***
GRADER    24   806.2   33.59  4.6828 2.694e-08 ***
Residuals 96   688.6    7.17                      
```

- The two ANOVA tables give identical SSE but different SS for `EXAM` and `GRADER`.
- Note the SS_{exam}, SS_{grader}, and SSE add up to the same number (SS_{total}) in both ANOVA tables as they are Type I ANOVA tables.
- Which ANOV table should we look at to determine the significance of treatment (GRADER)?

Model diagnostic for $y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$



How to Adjust Scores as Graders Were Inconsistent?

Based on the model $y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$, the score of the ith writing sample is $\mu + \beta_j$, which is estimated by $\widehat{\mu} + \widehat{\beta}_j$.

How to get $\widehat{\beta}_j$ in R? Recall R by default estimates parameters using the baseline constraints $\alpha_1 = \beta_1 = 0$, not the zero-sum constraints $\sum_{i=1}^{g} \alpha_i = \sum_{j=1}^{b} \beta_j = 0$.

One can use `constrasts()` and `contr.sum()` to force R using the the zero-sum constraints.

```
> contrasts(pr14.3$EXAM) = contr.sum(30)
> contrasts(pr14.3$GRADER) = contr.sum(25)
> lm1 = lm(score ~ EXAM + GRADER, data=pr14.3); lm1$coef
(Intercept)       EXAM1       EXAM2   .... (omitted)
     69.960     -12.568      -3.368
     EXAM28      EXAM29     GRADER1   .... (omotted)   GRADER24
     -2.128      16.192      -0.840                     0.160
```

Why is there no estimate for exam #30, nor for grader #25?

```
> muhat = lm1$coef[1]
> betahat = vector("numeric",length=30)
> betahat[1:29] = lm1$coef[2:30]
> betahat[30] = -sum(betahat[1:29])
> adjustedscore = muhat + betahat; adjustedscore
 [1] 57.392 66.592 84.392 75.152 69.472 56.376 51.616 60.416
 [9] 77.496 71.496 77.848 65.648 49.328 68.208 80.568 65.792
[17] 74.792 73.952 78.112 83.352 66.120 83.440 80.240 78.760
[25] 60.240 69.512 67.672 67.832 86.152 50.832

> adjustedscore = array(adjustedscore,dim=c(1,30))
> adjustedscore = data.frame(adjustedscore)
> colnames(adjustedscore) = 1:30
> adjustedscore
     1      2      3      4      5      6      7      8      9     10
57.392 66.592 84.392 75.152 69.472 56.376 51.616 60.416 77.496 71.496
    11     12     13     14     15     16     17     18     19     20
77.848 65.648 49.328 68.208 80.568 65.792 74.792 73.952 78.112 83.352
   21    22    23    24    25     26     27     28     29     30
66.12 83.44 80.24 78.76 60.24 69.512 67.672 67.832 86.152 50.832
```

Compare the adjusted score with the unadjusted scores (average of the 5 raw scores per exam).

```
> library(mosaic)
> unadjustedscore = mean(score ~ EXAM, data=pr14.3)
> unadjustedscore
   1    2    3    4    5    6    7    8    9   10   11   12   13   14   15 
57.4 66.0 84.6 75.2 69.8 55.4 53.4 59.0 78.4 71.2 77.8 65.4 49.2 69.4 79.8 
  16   17   18   19   20   21   22   23   24   25   26   27   28   29   30 
66.0 75.2 72.4 79.4 83.0 65.6 84.6 79.0 80.6 59.0 70.0 67.8 65.8 88.4 50.0 
```

Difference of the unadjusted and adjusted scores:

```
> unadjustedscore - adjustedscore
      1      2     3     4     5      6     7      8     9     10     11
1 0.008 -0.592 0.208 0.048 0.328 -0.976 1.784 -1.416 0.904 -0.296 -0.048
      12     13    14     15    16    17     18    19     20    21   22
1 -0.248 -0.128 1.192 -0.768 0.208 0.408 -1.552 1.288 -0.352 -0.52 1.16
     23   24    25    26    27     28    29     30
1 -1.24 1.84 -1.24 0.488 0.128 -2.032 2.248 -0.832
```

Exam #28 and #29 are off by over 2 points after adjustment.

How to Identify Inconsistent Graders?

We can do pairwise comparisons for the grader effects $\alpha_{i_1} - \alpha_{i_2}$ using the t-statistic $= \dfrac{\widehat{\alpha}_{i_1} - \widehat{\alpha}_{i_2}}{\text{SE}}$ where

$$SE = \sqrt{\text{MSE}\left(\frac{2k}{\lambda g}\right)} = \sqrt{7.17\left(\frac{2 \times 5}{1 \times 25}\right)} \approx 1.6935$$

with df $=$ (df of MSE) $= 96$.

The t-statistic must be $> t_{0.025,96} \approx 1.985$ to be significant at 5% level.

That is, $|\widehat{\alpha}_{i_1} - \widehat{\alpha}_{i_2}|$ must be at least

$$\text{LSD} = t_{0.025,96} \times SE \approx 1.985 \times 1.6935 \approx 3.362$$

to be significant at 5% level.

We have obtained $\widehat{\alpha}_1, \widehat{\alpha}_2, \ldots, \widehat{\alpha}_{24}$ in R on page 21.

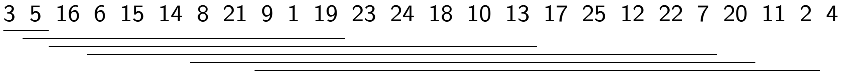
```
> lm1$coef[31:54]
 GRADER1  GRADER2  GRADER3  GRADER4  GRADER5  GRADER6  GRADER7  GRADER8
   -0.84     3.24    -6.36     7.48    -3.48    -2.36     1.60    -1.56
 GRADER9 GRADER10 GRADER11 GRADER12 GRADER13 GRADER14 GRADER15 GRADER16
   -1.12     0.48     2.16     1.32     0.76    -1.60    -1.60    -2.60
GRADER17 GRADER18 GRADER19 GRADER20 GRADER21 GRADER22 GRADER23 GRADER24
    1.24     0.20    -0.40     1.80    -1.24     1.52    -0.12     0.16
```

The last one can be computed as $\widehat{\alpha}_{25} = -\sum_{i=1}^{24} \widehat{\alpha}_i = 1.32$ as $\sum_{i=1}^{25} \widehat{\alpha}_i = 0$.

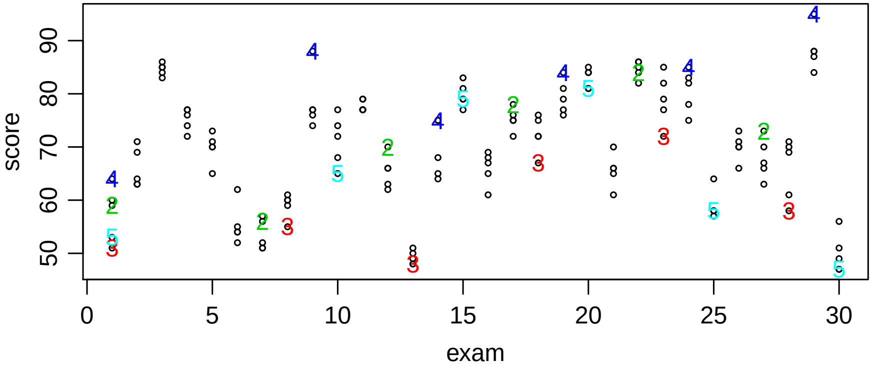
```
> alphahat25 = -sum(lm1$coef[31:54]); alphahat25
[1] 1.32
> names(alphahat25) = "GRADER25"
> alphahat = c(lm1$coef[31:54], alphahat25)
> alphahat
 GRADER1  GRADER2  GRADER3  GRADER4  GRADER5  GRADER6  GRADER7  GRADER8
   -0.84     3.24    -6.36     7.48    -3.48    -2.36     1.60    -1.56
 GRADER9 GRADER10 GRADER11 GRADER12 GRADER13 GRADER14 GRADER15 GRADER16
   -1.12     0.48     2.16     1.32     0.76    -1.60    -1.60    -2.60
GRADER17 GRADER18 GRADER19 GRADER20 GRADER21 GRADER22 GRADER23 GRADER24
    1.24     0.20    -0.40     1.80    -1.24     1.52    -0.12     0.16
GRADER25
    1.32
```

```
> sort(alphahat)
 GRADER3  GRADER5 GRADER16  GRADER6 GRADER15 GRADER14  GRADER8 GRADER21
   -6.36    -3.48    -2.60    -2.36    -1.60    -1.60    -1.56    -1.24
 GRADER9  GRADER1 GRADER19 GRADER23 GRADER24 GRADER18 GRADER10 GRADER13
   -1.12    -0.84    -0.40    -0.12     0.16     0.20     0.48     0.76
GRADER17 GRADER25 GRADER12 GRADER22  GRADER7 GRADER20 GRADER11  GRADER2
    1.24     1.32     1.32     1.52     1.60     1.80     2.16     3.24
 GRADER4
    7.48
```

Underline Diagram for pairwise comparison between graders:
(at 5% significance level, LSD = 3.362)



Grader #2, #3, #4, and #5 have their $\widehat{\alpha}_i$'s differ from most of other $\widehat{\alpha}_i$'s by more than 3.362. They appeared to be inconsistent with other graders.



- Grader #3 always gave the lowest score among the 5 graders grading the same exam
- Grader #4 always gave scores that substantially higher than the scores given by the other graders for the same exam.
- Grader #2 tends to give higher scores, Grader #5 tended to give lower scores, but not as much as Grader #3 and #4.