Chapter 5  Multiple Comparisons

Yibi Huang

- Why Worry About Multiple Comparisons?
- Familywise Error Rate
- Simultaneous Confidence Intervals
- Bonferroni’s Method
- Tukey-Kramer Procedure for Pairwise Comparisons
- Scheffé’s Method for Comparing All Contrasts
Why Worry About Multiple Comparisons?

Recall that, at level $\alpha = 0.05$, a hypothesis test will make a Type I error 5% of the time

- Type I error = $H_0$ being falsely rejected when it is true

What if we conduct multiple hypothesis tests?

- When 100 $H_0$’s are tested at 0.05 level, even if all $H_0$’s are true, it’s normal to have 5 being rejected.

- When multiple tests are done, it’s very likely that some significant results may be NOT be TRUE FINDINGS. The significance must be adjusted
Why Worry About Multiple Comparisons?

In an experiment, when the ANOVA F-test is rejected, we will attempt to compare ALL pairs of treatments, as well as contrasts to find treatments that are different from others.

For an experiment with $g$ treatments, there are

$$\binom{g}{2} = \frac{g(g-1)}{2}$$ pairwise comparisons to make, and numerous contrasts.

When many $H_0$’s are tested, it’s very likely that some of them are falsely rejected even if all of $H_0$’s are true as we would falsely reject every true $H_0$ at 5% level about 5% of the time.
10 groups of observations of size 5 each are generated from the $N(0,1)$ distribution.

```r
g = 10  # number of treatments
n = 5   # number of replicates per treatment

trt = gl(g, n, labels=LETTERS[1:g])  # Treatment: A, B, C, ..., I, J
y = rnorm(g*n, mean=0, sd = 1)        # Standard normal

data.frame(trt,y)
```

The data looks like

```
    trt y
1   A -1.133072151
2   A -1.155419923
3   A  0.287352711
4   A -0.095260234
5   A -0.530695825
6   B  0.815546733
7   B  0.694283605
(...omitted...)
49  J -0.741629226
50  J -1.355834197
```

Chapter 5 - 4
As all the y’s are generated from the $N(0,1)$ distribution, no pair of treatments should be significantly different, but . . .

```r
> pairwise.t.test(y, trt, p.adjust.method = "none")
```

Pairwise comparisons using t tests with pooled SD

data:  y and trt

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>0.816</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>C</td>
<td>0.328</td>
<td>0.454</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>D</td>
<td>0.250</td>
<td>0.356</td>
<td>0.860</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>E</td>
<td>0.206</td>
<td>0.300</td>
<td>0.769</td>
<td>0.907</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>F</td>
<td>0.267</td>
<td>0.377</td>
<td>0.891</td>
<td>0.968</td>
<td>0.876</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>G</td>
<td>0.656</td>
<td>0.831</td>
<td>0.592</td>
<td>0.477</td>
<td>0.408</td>
<td>0.502</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>H</td>
<td>0.039</td>
<td>0.065</td>
<td>0.260</td>
<td>0.341</td>
<td>0.402</td>
<td>0.321</td>
<td>0.100</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>I</td>
<td>0.066</td>
<td>0.106</td>
<td>0.374</td>
<td>0.475</td>
<td>0.550</td>
<td>0.451</td>
<td>0.158</td>
<td>0.809</td>
<td>-</td>
</tr>
<tr>
<td>J</td>
<td>0.565</td>
<td>0.731</td>
<td>0.685</td>
<td>0.561</td>
<td>0.485</td>
<td>0.588</td>
<td>0.896</td>
<td>0.129</td>
<td>0.198</td>
</tr>
</tbody>
</table>

P value adjustment method: none
Repeat the following several times.

g = 10       # number of treatments
n = 5        # number of replicates per treatment
trt = gl(g, n, labels=LETTERS[1:g])  # Treatment: A, B, C, ..., I, J
y = rnorm(g*n, mean=0, sd = 1)       # Standard normal
pairwise.t.test(y, trt, p.adjust.method = "none")

How often do you see a significant difference?
Data Snooping

- If one looks at data first and decide which contrast to test based on what they see, that is called \textit{data snooping}, e.g.,
  - when one decides to compare treatment \( A \) & \( E \) because \( A \) has the highest mean and \( E \) the lowest
  - or if one decides to test the contrast

\[
C = \frac{\mu_A + \mu_C}{2} - \frac{\mu_B + \mu_D}{2}
\]

because \( A \) and \( C \) have higher means than \( B \) and \( D \)

- Data snooping is problematic because when people choose the pair of treatments with the greatest difference or contrast with a big effect after looking at data, they have implicitly tested many pairs and contrasts that are unlikely to be significant. Effectively, they have conducted many tests. They cannot pretend as if they’ve just done one.

- If a comparison or contrast is determined after looking at the data (data snooping), one must adjust for multiple comparisons.
5.1 Familywise Error Rate (FWER)

Given a single null hypothesis $H_0$,

- recall a *Type I error* occurs when $H_0$ is true but is rejected;
- the *level* (or *size*, or *Type I error rate*) of a test is the chance of making a Type I error.

Given a family of null hypotheses $H_{01}, H_{02}, \ldots, H_{0k}$,

- a *familywise Type I error* occurs if $H_{01}, H_{02}, \ldots, H_{0k}$ are all true but at least one of them is rejected;
- The *familywise error rate (FWER)*, also called *experimentwise error rate*, is defined as the chance of making a familywise Type I error

$$FWER = P(\text{at least one of } H_{01}, \ldots, H_{0k} \text{ is falsely rejected})$$

- FWER depends on the *family*. The larger the family, the larger the FWER.
Simultaneous Confidence Intervals

Similarly, a level 95% confidence level \((L, U)\) for a parameter \(\theta\) may fail to cover \(\theta\) 5% of the time.

What if we construct multiple 95% confidence intervals

\[
\{(L_1, U_1), (L_2, U_2), \ldots, (L_k, U_k)\}
\]

for several different parameters \(\theta_1, \theta_2, \ldots, \theta_k\), the chance that at least one of the intervals fails to cover the parameter is (a lot) more than 5%.
Simultaneous Confidence Intervals

Given a family of parameters \( \{ \theta_1, \theta_2, \ldots, \theta_k \} \), a 100(1 − \( \alpha \))% simultaneous confidence intervals is a family of intervals

\[
\{(L_1, U_1), (L_2, U_2), \ldots, (L_k, U_k)\}
\]

that

\[
P(L_i \leq \theta_i \leq U_i \text{ for all } i) > 1 - \alpha.
\]

Note here that \( L_i \)'s and \( U_i \)'s are random variables that depends on the data.
Multiple Comparisons

To account for multiple comparisons, we will need to make our C.I. wider, and the critical values larger to ensure the chance of making any false rejection $< \alpha$.

We will introduce several multiple comparison methods. All of them produce simultaneous C.I.’s of the form

$$\text{estimate} \pm (\text{critical value}) \times (\text{SE of the estimate})$$

and reject $H_0$ when

$$|t_0| = \frac{|\text{estimate}|}{\text{SE of the estimate}} > \text{critical value}.$$ 

Here the “estimates” and “SEs” are identical to those in the usual $t$-tests and $t$-intervals. Only the critical values change with the adjustment methods, as summarized on Slide 32.
5.2 Bonferroni’s Method

Given that $H_{01}, \ldots, H_{0k}$ being all true, by the Bonferroni’s inequality we know

$$\text{FWER} = P(\text{at least one of } H_{01}, \ldots, H_{0k} \text{ is rejected})$$

$$\leq \sum_{i=1}^{k} P(H_{0i} \text{ is rejected})$$

**(type I error rate for $H_{0i}$)**

If the Type I error rate for each of the $k$ nulls can be controlled at $\alpha/k$, then

$$\text{FWER} \leq \sum_{i=1}^{k} \frac{\alpha}{k} = \alpha.$$ 

Bonferroni’s method rejects a null if

the comparisonwise $P$-value is less than $\alpha/k$, 

or equivalently if

the $t$-statistic $> t_{df, \alpha/2/k}$. 

Chapter 5 - 12
Example: Grass/Weed Competition (Ch 3)

Big bluestem was first seeded in these plots. One year later, quack grass was seeded to each plot.

**Response**: Percentage of living material in each plot that is big bluestem one year after quack grass was seeded.

<table>
<thead>
<tr>
<th>Treatment</th>
<th>1N</th>
<th>1Y</th>
<th>2N</th>
<th>3N</th>
<th>4N</th>
<th>4Y</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>97</td>
<td>83</td>
<td>85</td>
<td>64</td>
<td>52</td>
<td>48</td>
</tr>
<tr>
<td></td>
<td>96</td>
<td>87</td>
<td>84</td>
<td>72</td>
<td>56</td>
<td>58</td>
</tr>
<tr>
<td></td>
<td>92</td>
<td>78</td>
<td>78</td>
<td>63</td>
<td>44</td>
<td>49</td>
</tr>
<tr>
<td></td>
<td>95</td>
<td>81</td>
<td>79</td>
<td>74</td>
<td>50</td>
<td>53</td>
</tr>
</tbody>
</table>

```
grass = read.table("http://www.stat.uchicago.edu/~yibi/s222/grassweed.txt", h=T)
```
Example: Grass/Weed Competition (Ch 3)

The group means of the 4 treatments are

```r
> library(mosaic)
> mean(percent ~ trt, data = grass)
    1N  1Y  2N  3N  4N  4Y
  95.00 82.25 81.50 68.25 50.50 52.00
```

From the ANOVA table below, we get $MSE = 17.97$.

```r
> lm1 = lm(percent ~ trt, data = grass)
> anova(lm1)
Response: percent
   Df Sum Sq Mean Sq F value  Pr(>F)
trt   5 6398.3 1279.67  71.203 3.197e-11 ***
Residuals 18 323.5  17.97
```

The SE for pairwise comparison is

$$SE = \sqrt{MSE \left( \frac{1}{n_i} + \frac{1}{n_j} \right)} = \sqrt{17.97 \left( \frac{1}{4} + \frac{1}{4} \right)} \approx 2.9975$$
Example — Grass/Weed (Bonferoni’s Method)

To be significant at FWER = \( \alpha \) based on Bonferoni’s correction, the \( t \)-statistic for pairwise comparison must be at least

\[
t = \frac{\overline{y}_i - \overline{y}_j}{SE} > t_{N-g,\alpha/2/k}
\]

where \( k = 15 \) since there are \((g^2) = \binom{6}{2} = 15\) pairs to compare.

\[
df = N - g = 24 - 6 = 18, \quad t_{N-g,\alpha/2/k} = t_{18,0.05/2/15} \approx 3.38.
\]

\[
> qt(0.05/2/15, df=18, lower.tail=F)
\]

[1] 3.380362

So a pair of treatments \( i, j \) are significantly different at FWER = 0.05 iff

\[
|\overline{y}_i - \overline{y}_j| > SE \times t_{N-g,\alpha/2/k} \approx 2.9975 \times 3.38 \approx 10.13 = \text{BSD}.
\]

This is called **Bonferoni’s Significant Difference (BSD)**.

\[
\begin{array}{ccccccc}
4N & 4Y & 3N & 2N & 1Y & 1N \\
\hline
50.50 & 52.00 & 68.25 & 81.50 & 82.25 & 95.00
\end{array}
\]
Alternatively, one can compute the $P$-values based on the ordinary pairwise $t$-test

```r
> pairwise.t.test(grass$percent, grass$trt, p.adjust="none")
Pairwise comparisons using t tests with pooled SD
data: grass$percent and grass$trt

1N  1Y  2N  3N  4N
1Y  0.00048 -  -  -  -
2N  0.00027 0.80527 -  -  -
3N  5.0e-08 0.00019 0.00033 -  -
4N  1.5e-11 3.7e-09 5.3e-09 1.3e-05 -
4Y  2.7e-11 7.8e-09 1.1e-08 3.8e-05 0.62287

P value adjustment method: none

There are $k = \binom{6}{2} = 15$ tests.

To keep FWER $\leq \alpha = 0.05$, instead of rejecting a null when the $P$-value $< \alpha$, Bonferoni’s method rejects when

$$\text{the } P\text{-value } < \frac{\alpha}{k} = \frac{0.05}{15} \approx 0.0033.$$ 

Only (1Y, 2N) and (4N, 4Y) are insignificant.
> pairwise.t.test(grass$percent, grass$trt,  
  p.adjust.method = "bonferroni")

Pairwise comparisons using t tests with pooled SD

data: grass$percent and grass$trt

<table>
<thead>
<tr>
<th></th>
<th>1N</th>
<th>1Y</th>
<th>2N</th>
<th>3N</th>
<th>4N</th>
</tr>
</thead>
<tbody>
<tr>
<td>1N</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>1Y</td>
<td>0.00717</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>2N</td>
<td>0.00412</td>
<td>1.00000</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>3N</td>
<td>7.5e-07</td>
<td>0.00286</td>
<td>0.00496</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>4N</td>
<td>2.3e-10</td>
<td>5.5e-08</td>
<td>8.0e-08</td>
<td>0.00020</td>
<td>-</td>
</tr>
<tr>
<td>4Y</td>
<td>4.1e-10</td>
<td>1.2e-07</td>
<td>1.7e-07</td>
<td>0.00057</td>
<td>1.00000</td>
</tr>
</tbody>
</table>

P value adjustment method: bonferroni

Each Bonferoni $P$-value is the corresponding unadjusted $P$-value multiplied $k$.
You can just compare the Bonferroni $P$-values with $\alpha$.  

Chapter 5 - 17
Limitation of Bonferroni’s Method

- The number of tests $k$ must be finite.
- Bonferroni’s method works OK when the number of tests $k$ is small.
- When the number of tests $k$ is large ($>10$), Bonferroni often get too conservative (too hard to reject $H_0$) than necessary. The actual FWER can be much less than $\alpha$. 
5.4 Tukey-Kramer Procedure for Pairwise Comparisons

- Family: ALL PAIRWISE COMPARISON $\mu_i - \mu_k$

- For a balanced design ($n_1 = \ldots = n_g = n$), observe that

$$|t_0| = \frac{|\bar{y}_{i\cdot} - \bar{y}_{k\cdot}|}{\sqrt{\text{MSE} \left( \frac{1}{n} + \frac{1}{n} \right)}} \leq \frac{\bar{y}_{\text{max}} - \bar{y}_{\text{min}}}{\sqrt{2\text{MSE}/n}} = \frac{q}{\sqrt{2}}.$$ 

in which $q = \frac{\bar{y}_{\text{max}} - \bar{y}_{\text{min}}}{\sqrt{\text{MSE}/n}}$ has a studentized range distribution.

- The critical values $q_\alpha(g, N - g)$ for the studentized range distribution can be found on p.633-634, Table D.8 in the textbook.

- Controls the (strong) FWER exactly at $\alpha$ for balanced designs ($n_1 = \ldots = n_g$); approximately at $\alpha$ for unbalanced designs.
Tukey-Kramer Procedure for All Pairwise Comparisons

For all $1 \leq i \neq k \leq g$, the $100(1 - \alpha)\%$ Tukey-Kramer’s simultaneous C.I. for $\mu_i - \mu_k$ is

$$
\bar{y}_i - \bar{y}_k \pm \frac{q_\alpha(g, N - g)}{\sqrt{2}} \text{SE}(\bar{y}_i - \bar{y}_k)
$$

For $H_0 : \mu_i - \mu_k = 0$ v.s. $H_a : \mu_i - \mu_k \neq 0$, reject $H_0$ if

$$
|t_0| = \frac{|\bar{y}_i - \bar{y}_k|}{\text{SE}(\bar{y}_i - \bar{y}_k)} > \frac{q_\alpha(g, N - g)}{\sqrt{2}}
$$

In both the C.I. and the test,

$$
\text{SE}(\bar{y}_i - \bar{y}_k) = \sqrt{\text{MSE} \left( \frac{1}{n_i} + \frac{1}{n_k} \right)}.
$$
Tukey’s HSD

For a balanced design \((n_1 = \ldots = n_g = n)\), to be significant at FWER = \(\alpha\) based Tukey’s correction, the mean difference between a pair of treatments must be at least

\[
\frac{q_\alpha(g, N-g)}{\sqrt{2}} \times \sqrt{\text{MSE} \left( \frac{1}{n} + \frac{1}{n} \right)}
\]

This is called Tukey’s Honest Significant Difference (Tukey’s HSD).

R command to find \(q_\alpha(a, f)\): \(\text{qtukey}(1-\alpha, g, N-g)\)

\>

> qtukey(0.95, 6, 18)/sqrt(2)

[1] 3.178035

For the Grass/Weed example, Tukey’s HSD is

\[
3.178 \times \sqrt{17.97 \left( \frac{1}{4} + \frac{1}{4} \right)} \approx 9.526
\]

<table>
<thead>
<tr>
<th>4N</th>
<th>4Y</th>
<th>3N</th>
<th>2N</th>
<th>1Y</th>
<th>1N</th>
</tr>
</thead>
<tbody>
<tr>
<td>50.50</td>
<td>52.00</td>
<td>68.25</td>
<td>81.50</td>
<td>82.25</td>
<td>95.00</td>
</tr>
</tbody>
</table>

Chapter 5 - 21
Tukey’s HSD in R

TukeyHSD command only works for `aov()` model, not `lm()` model.

```r
> aov1 = aov(percent ~ trt, data = grass)
> TukeyHSD(aov1)

Tukey multiple comparisons of means
95% family-wise confidence level
Fit: aov(formula = percent ~ trt, data = grass)
$trt

<table>
<thead>
<tr>
<th>diff</th>
<th>lwr</th>
<th>upr</th>
<th>p adj</th>
</tr>
</thead>
<tbody>
<tr>
<td>1Y-1N</td>
<td>-12.75</td>
<td>-22.276745</td>
<td>-3.223255</td>
</tr>
<tr>
<td>2N-1N</td>
<td>-13.50</td>
<td>-23.026745</td>
<td>-3.973255</td>
</tr>
<tr>
<td>3N-1N</td>
<td>-26.75</td>
<td>-36.276745</td>
<td>-17.223255</td>
</tr>
<tr>
<td>4N-1N</td>
<td>-44.50</td>
<td>-54.026745</td>
<td>-34.973255</td>
</tr>
<tr>
<td>4Y-1N</td>
<td>-43.00</td>
<td>-52.526745</td>
<td>-33.473255</td>
</tr>
<tr>
<td>2N-1Y</td>
<td>-0.75</td>
<td>-10.276745</td>
<td>8.776745</td>
</tr>
<tr>
<td>3N-1Y</td>
<td>-14.00</td>
<td>-23.526745</td>
<td>-4.473255</td>
</tr>
<tr>
<td>4N-1Y</td>
<td>-31.75</td>
<td>-41.276745</td>
<td>-22.223255</td>
</tr>
<tr>
<td>4Y-1Y</td>
<td>-30.25</td>
<td>-39.776745</td>
<td>-20.723255</td>
</tr>
<tr>
<td>3N-2N</td>
<td>-13.25</td>
<td>-22.776745</td>
<td>-3.723255</td>
</tr>
<tr>
<td>4N-2N</td>
<td>-31.00</td>
<td>-40.526745</td>
<td>-21.473255</td>
</tr>
<tr>
<td>4Y-2N</td>
<td>-29.50</td>
<td>-39.026745</td>
<td>-19.973255</td>
</tr>
<tr>
<td>4N-3N</td>
<td>-17.75</td>
<td>-27.276745</td>
<td>-8.223255</td>
</tr>
<tr>
<td>4Y-3N</td>
<td>-16.25</td>
<td>-25.776745</td>
<td>-6.723255</td>
</tr>
</tbody>
</table>
```
Tukey’s HSD in R

> TukeyHSD(aov1)
  Tukey multiple comparisons of means
  95% family-wise confidence level

$trt
  diff    lwr     upr     p adj
1Y-1N  -12.75  -22.277  -3.223  0.0054014
2N-1N  -13.50  -23.027  -3.973  0.0031780
3N-1N  -26.75  -36.277  -17.223 0.0000007
4N-1N  -44.50  -54.027  -34.973 0.0000000
...(omitted)...

Note that the widths of all CIs above are 2x of the HSD. E.g., the width of the CI for 1Y-1N is

\[-3.223255 - (-22.276745) = 19.05349\]

is twice of HSD \(\approx 9.526\).
5.3 Scheffé’s Method for Comparing All Contrasts

Suppose there are $g$ treatments in total. Consider a contrast $C = \sum_{i=1}^{g} \omega_i \mu_i$. Recall

$$\hat{C} = \sum_{i=1}^{g} \omega_i \bar{y}_i \cdot, \quad SE(\hat{C}) = \sqrt{\text{MSE} \times \sum_{i=1}^{g} \frac{\omega_i^2}{n_i}}$$

The $100(1 - \alpha)\%$ Scheffé’s simultaneous C.I. for all contrasts $C$ is

$$\hat{C} \pm \sqrt{(g - 1)F_{\alpha, g-1, N-g}}SE(\hat{C})$$

For testing $H_0 : C = 0$ v.s. $H_a : C \neq 0$, reject $H_0$ when

$$|t_0| = \frac{|\hat{C}|}{SE(\hat{C})} > \sqrt{(g - 1)F_{\alpha, g-1, N-g}}$$
Scheffé’s Method for Comparing All Contrasts

- Most conservative (least powerful) of all tests. *Protects against data snooping!*

- Controls (strong) FWER at $\alpha$, where the family is ALL POSSIBLE CONTRASTS

- Should be used if you have not planned contrasts in advance.
Proof of Scheffé’s Method (1)

Because \( \sum_{i=1}^{g} \omega_i = 0 \), observe that

\[
\hat{C} = \sum_{i=1}^{g} \omega_i \bar{y}_{i\bullet} = \sum_{i=1}^{g} \omega_i (\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet}).
\]

By the Cauchy-Schwartz Inequality

\[
|\sum a_i b_i| \leq \sqrt{\sum a_i^2 \sum b_i^2}
\]

and let \( a_i = \frac{\omega_i}{\sqrt{n_i}} \) and \( b_i = \sqrt{n_i} (\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet}) \), we get

\[
|\hat{C}| = \left| \sum_{i=1}^{g} \omega_i (\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet}) \right| \leq \sqrt{\sum_{i=1}^{g} \frac{\omega_i^2}{n_i} \sum_{i=1}^{g} n_i (\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet})^2}
\]

Recall that \( SS_{trt} = \sum_{i=1}^{g} n_i (\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet})^2 \), we get the inequality

\[
|\hat{C}| \leq \sqrt{\sum_{i=1}^{g} \frac{\omega_i^2}{n_i} SS_{trt}}.
\]

Chapter 5 - 26
Proof of Scheffé’s Method (2)

Recall the $t$-statistic for testing $H_0: C = 0$ is $t_0(C) = \frac{\hat{C}}{SE(\hat{C})}$, and using the inequality $|\hat{C}| \leq \sqrt{\sum_{i=1}^{g} \frac{\omega_i^2}{n_i} SS_{trt}}$ proved in the previous page, we have

$$|t_0(C)| = \frac{|\hat{C}|}{SE(\hat{C})} = \frac{|\hat{C}|}{\sqrt{MSE \sum_{i=1}^{g} \frac{\omega_i^2}{n_i}}} \leq \frac{\sqrt{\sum_{i=1}^{g} \frac{\omega_i^2}{n_i} SS_{trt}}}{\sqrt{MSE \sum_{i=1}^{g} \frac{\omega_i^2}{n_i}}} = \sqrt{\frac{SS_{trt}}{MSE}}$$

Recall $F = \frac{MS_{trt}}{MSE}$ is the ANOVA $F$-statistic, we have

$$|t_0(C)| \leq \sqrt{\frac{SS_{trt}}{MSE}} = \sqrt{\frac{(g - 1)MS_{trt}}{MSE}} = \sqrt{(g - 1)F}.$$ 

We thus get a uniform upper bound for the $t$-statistic for any contrast $C$

$$|t_0(C)| \leq \sqrt{(g - 1)F}.$$
Recall that $F$ has a $F$-distribution with $g - 1$ and $N - g$ degrees of freedom, so $P(F > F_{\alpha, g-1, N-g}) = \alpha$.

Since $|t_0(C)| < \sqrt{(g - 1)F}$, we can see that

$$FWER = P \left( |t_0(C)| > \sqrt{(g - 1)F_{\alpha, g-1, N-g}} \text{ for any contrast } C \right)$$

$$\leq P \left( \sqrt{(g - 1)F} > \sqrt{(g - 1)F_{\alpha, g-1, N-g}} \right)$$

$$= P(F > F_{\alpha, g-1, N-g}) = \alpha.$$
A Contrast for Nitrogen Effect

<table>
<thead>
<tr>
<th>Group</th>
<th>1N</th>
<th>1Y</th>
<th>2N</th>
<th>3N</th>
<th>4N</th>
<th>4Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \bar{y}_{i.} )</td>
<td>95</td>
<td>82.25</td>
<td>81.5</td>
<td>68.25</td>
<td>50.5</td>
<td>52</td>
</tr>
</tbody>
</table>

The contrast we consider is

\[
C = \frac{\mu_{1N} + \mu_{1Y}}{2} - \frac{\mu_{4N} + \mu_{4Y}}{2}
\]

which is estimated to be

\[
\hat{C} = \frac{\bar{y}_{1N} + \bar{y}_{1Y}}{2} - \frac{\bar{y}_{4N} + \bar{y}_{4Y}}{2} = \frac{95 + 82.25}{2} - \frac{50.5 + 52}{2} = 37.375.
\]

with the standard error

\[
SE(\hat{C}) = \sqrt{\text{MSE} \sum_{i=1}^{g} \frac{\omega_{i}^2}{n_i}} = \sqrt{17.97 \left( \frac{0.5^2}{4} + \frac{0.5^2}{4} + \frac{(-0.5)^2}{4} + \frac{(-0.5)^2}{4} \right)} \approx 2.12.
\]

To test \( H_0: C = 0 \) v.s. \( H_a: C \neq 0 \), the \( t \)-statistic is

\[
t = \frac{\hat{C}}{SE(\hat{C})} \approx \frac{37.375}{2.12} \approx 17.63.
\]
A Contrast for Nitrogen Effect

With Scheffé Method, the critical value controlling FWER at 0.05 is

$$\sqrt{(g - 1)F_{\alpha, g-1, N-g}} = \sqrt{(6 - 1)F_{0.05, 6-1, 24-6}} \approx \sqrt{(6 - 1) \times 2.77} \approx 3.72$$

> qf(0.05, df1=6-1, df2=24-6, lower.tail=F)
[1] 2.772853
> sqrt((6-1)*qf(0.05, df1=6-1, df2=24-6, lower.tail=F))
[1] 3.723475

The critical value 3.72 for Scheffé’s method means that: if all treatments are equal, the contrast with the greatest $t$-statistic will exceed 3.72 for only 5% of the time. The magnitude of the $t$-statistic 17.63 for the contrast we considered is far above the critical value 3.72.

Conclusion: We can be certain that the contrast is really significant, even if the contrast was suggested by data snooping.
5.4.7 Fisher’s Least Significant Difference (LSD)

- The **least significant difference** (LSD) is the minimum amount by which two means must differ in order to be considered statistically different.

- LSD = the usual $t$-tests and $t$-intervals
  NO adjustment is made for multiple comparisons

- **least conservative** (most likely to reject) among all procedures, FWER can be large when family of tests is large

- too liberal, but greater power (more likely to reject)
## Summary of Multiple Comparison Adjustments

<table>
<thead>
<tr>
<th>Method</th>
<th>Family of Tests</th>
<th>Critical Value to Keep FWER $&lt; \alpha$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fisher’s LSD</td>
<td>a single pairwise comparison</td>
<td>$t_{\alpha/2,N-g}$</td>
</tr>
<tr>
<td>Tukey-Kramer</td>
<td>all pairwise comparisons</td>
<td>$q_{\alpha}(g, N - g)/\sqrt{2}$</td>
</tr>
<tr>
<td>Bonferroni</td>
<td>varies</td>
<td>$t_{\alpha/(2k),N-g}$, where $k = # \text{ of tests}$</td>
</tr>
<tr>
<td>Scheffé</td>
<td>all contrasts</td>
<td>$\sqrt{(g - 1)F_{\alpha,g-1,N-g}}$</td>
</tr>
</tbody>
</table>
Which Procedures to Use?

▶ Use BONFERRONI when only interested in a small number of planned contrasts (or pairwise comparisons)

▶ Use TUKEY when only interested in all (or most) pairwise comparisons of means

▶ Use SCHEFFE when doing anything that could be considered data snooping – i.e. for any unplanned contrasts
In the figure above, Bonferroni is the Bonferroni for all pairwise comparisons.

For a smaller family of, say $k$ tests, one can divide $\alpha$ by $k$ rather than by $r = \frac{g(g-1)}{2}$. The resulting C.I. or tests may have stronger power than Tukey or Dunnett, will keeping FWER $< \alpha$.

Remember to use Bonferroni the contrasts should be pre-planned.
Multiple Comparisons in Balanced Block Designs

All the multiple comparison procedures apply to all balanced block designs just change the degree of freedom from $N - g$ to the d.f. of MSE

<table>
<thead>
<tr>
<th>Method</th>
<th>Family of Tests</th>
<th>Critical Value to Keep FWER $&lt; \alpha$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fisher’s LSD</td>
<td>a single pairwise comparison</td>
<td>$t_{\alpha/2, \text{df of MSE}}$</td>
</tr>
<tr>
<td>Tukey-Kramer</td>
<td>all pairwise comparisons</td>
<td>$q_{\alpha}(g, \text{df of MSE})/\sqrt{2}$</td>
</tr>
<tr>
<td>Bonferroni</td>
<td>all pairwise comparisons</td>
<td>$t_{\alpha/(2r), \text{df of MSE}}$, where $r = \frac{g(g-1)}{2}$</td>
</tr>
<tr>
<td>Scheffé</td>
<td>all contrasts</td>
<td>$\sqrt{(g-1)F_{\alpha,g-1, \text{df of MSE}}}$</td>
</tr>
</tbody>
</table>

Chapter 5 - 35
Recall Example 13.1 (Mealybugs on Cycads)

- Treatment: water (control), fungal spores, and horticultural oil
- 5 infested cycads, 3 branches are randomly chosen on each cycad, and 2 patches (3 cm × 3 cm) are marked on each branch
- 3 branches on each cycad are randomly assigned to the 3 treatments
- Response: difference of the # of mealybugs in the patches before and 3 days after treatments are applied
- As the patches are measurement units, we take the average of the two patches on each branch as the response

<table>
<thead>
<tr>
<th>Plant</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Water</td>
<td>-9</td>
<td>18</td>
<td>10</td>
<td>9</td>
<td>-6</td>
</tr>
<tr>
<td></td>
<td>-6</td>
<td>5</td>
<td>9</td>
<td>0</td>
<td>13</td>
</tr>
<tr>
<td>Spores</td>
<td>-4</td>
<td>29</td>
<td>4</td>
<td>-2</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>10</td>
<td>-1</td>
<td>6</td>
<td>-1</td>
</tr>
<tr>
<td>Oil</td>
<td>4</td>
<td>29</td>
<td>14</td>
<td>14</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>11</td>
<td>36</td>
<td>16</td>
<td>18</td>
<td>15</td>
</tr>
</tbody>
</table>
Example 13.1 (Mealybugs on Cycads)

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Water</th>
<th>Spore</th>
<th>Oil</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\bar{y}_i$</td>
<td>4.3</td>
<td>5.9</td>
<td>16.4</td>
</tr>
</tbody>
</table>

The SE for pairwise comparison is

$$\sqrt{\text{MSE} \left( \frac{1}{r} + \frac{1}{r} \right)} = \sqrt{17.725 \left( \frac{1}{5} + \frac{1}{5} \right)} \approx 2.663.$$  

Tukey’s critical value is 2.857.

> qtukey(0.95, 3, df = 8)/sqrt(2)

[1] 2.857444

Tukey’s HSD controlling FWER at 0.05 is 2.857 \times 2.663 \approx 7.608.

<table>
<thead>
<tr>
<th>Water</th>
<th>Spore</th>
<th>Oil</th>
</tr>
</thead>
</table>

We see that spores treatment cannot be distinguished from the control (water) (their mean did not differ by more than 7.608), but both can be distinguished from the oil treatment.
Example 13.1 (Mealybugs on Cycads)

```r
> aov1 = aov(avechange ~ trt + as.factor(plant), data=cycad)
> TukeyHSD(aov1)

Tukey multiple comparisons of means
95% family-wise confidence level

$trt$

<table>
<thead>
<tr>
<th>diff</th>
<th>lwr</th>
<th>upr</th>
<th>p adj</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spore-Water</td>
<td>1.6</td>
<td>-6.008532</td>
<td>9.208532</td>
</tr>
<tr>
<td>Oil-Water</td>
<td>12.1</td>
<td>0.01491468</td>
<td>19.708532</td>
</tr>
<tr>
<td>Oil-Spore</td>
<td>10.5</td>
<td>2.891468</td>
<td>18.108532</td>
</tr>
</tbody>
</table>

$'as.factor(plant)'$

<table>
<thead>
<tr>
<th>diff</th>
<th>lwr</th>
<th>upr</th>
<th>p adj</th>
</tr>
</thead>
<tbody>
<tr>
<td>2-1</td>
<td>20.666667</td>
<td>8.790833</td>
<td>32.5425005</td>
</tr>
<tr>
<td>3-1</td>
<td>8.166667</td>
<td>-3.709167</td>
<td>20.0425005</td>
</tr>
<tr>
<td>4-1</td>
<td>7.000000</td>
<td>-4.875834</td>
<td>18.8758339</td>
</tr>
<tr>
<td>5-1</td>
<td>6.000000</td>
<td>-5.875834</td>
<td>17.8758339</td>
</tr>
<tr>
<td>3-2</td>
<td>-12.500000</td>
<td>-24.375834</td>
<td>-0.6241661</td>
</tr>
<tr>
<td>4-2</td>
<td>-13.666667</td>
<td>-25.542501</td>
<td>-1.7908328</td>
</tr>
<tr>
<td>5-2</td>
<td>-14.666667</td>
<td>-26.542501</td>
<td>-2.7908328</td>
</tr>
<tr>
<td>4-3</td>
<td>-1.166667</td>
<td>-13.042501</td>
<td>10.7091672</td>
</tr>
<tr>
<td>5-3</td>
<td>-2.166667</td>
<td>-14.042501</td>
<td>9.7091672</td>
</tr>
<tr>
<td>5-4</td>
<td>-1.000000</td>
<td>-12.875834</td>
<td>10.8758339</td>
</tr>
</tbody>
</table>
```

Tukey’s HSD at 5% level for pairwise comparisons of the 3 treatments agrees with our computation.

Tukey’s HSD for pairwise comparisons of the 5 plants is nonsense here.
Example: Problem 13.4 in Oehlert’s Book (HW12)

- 2 replications of 5 × 5 Latin Squares
- two blocking factors: grader and student
- graders are reused but students are not
- treatment: exam

<table>
<thead>
<tr>
<th>Student</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>68 D</td>
<td>65 A</td>
<td>76 E</td>
<td>74 C</td>
<td>76 B</td>
</tr>
<tr>
<td>2</td>
<td>68 A</td>
<td>77 E</td>
<td>84 B</td>
<td>65 D</td>
<td>75 C</td>
</tr>
<tr>
<td>3</td>
<td>73 C</td>
<td>85 B</td>
<td>72 D</td>
<td>68 E</td>
<td>62 A</td>
</tr>
<tr>
<td>4</td>
<td>74 E</td>
<td>76 C</td>
<td>57 A</td>
<td>79 B</td>
<td>64 D</td>
</tr>
<tr>
<td>5</td>
<td>80 B</td>
<td>71 D</td>
<td>76 C</td>
<td>59 A</td>
<td>68 E</td>
</tr>
<tr>
<td>6</td>
<td>69 D</td>
<td>75 E</td>
<td>81 B</td>
<td>68 A</td>
<td>68 C</td>
</tr>
<tr>
<td>7</td>
<td>60 C</td>
<td>62 D</td>
<td>62 E</td>
<td>66 B</td>
<td>40 A</td>
</tr>
<tr>
<td>8</td>
<td>70 B</td>
<td>55 A</td>
<td>62 C</td>
<td>57 E</td>
<td>40 D</td>
</tr>
<tr>
<td>9</td>
<td>61 E</td>
<td>67 C</td>
<td>53 A</td>
<td>63 D</td>
<td>69 B</td>
</tr>
<tr>
<td>10</td>
<td>37 A</td>
<td>53 B</td>
<td>31 D</td>
<td>48 C</td>
<td>33 E</td>
</tr>
</tbody>
</table>

Model: \[ y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + \epsilon_{ijk} \]

(score) (exam) (student) (grader)

In HW12, we tested the contrast

\[ C = \alpha_A - \frac{\alpha_C + \alpha_D + \alpha_E}{3} \]
> mydata = read.table("http://users.stat.umn.edu/~gary/book/fcdae.data/pr13.4", h=T)
> lm1 = lm(score ~ as.factor(student)+as.factor(grader)+as.factor(exam),
    data=mydata)
> anova(lm1)

Response: score

<table>
<thead>
<tr>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>as.factor(student)</td>
<td>9</td>
<td>5050.3</td>
<td>561.15</td>
<td>40.3376</td>
</tr>
<tr>
<td>as.factor(grader)</td>
<td>4</td>
<td>443.3</td>
<td>110.83</td>
<td>7.9669</td>
</tr>
<tr>
<td>as.factor(exam)</td>
<td>4</td>
<td>1889.9</td>
<td>472.48</td>
<td>33.9639</td>
</tr>
<tr>
<td>Residuals</td>
<td>32</td>
<td>445.2</td>
<td>13.91</td>
<td></td>
</tr>
</tbody>
</table>

> library(mosaic)
> mean(score ~ exam, data=mydata)

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>56.4</td>
<td>74.3</td>
<td>67.9</td>
<td>60.5</td>
<td>65.1</td>
</tr>
</tbody>
</table>

\[ \hat{C} = \bar{y}_{\text{A..}} - \frac{\bar{y}_{\text{C..}} + \bar{y}_{\text{D..}} + \bar{y}_{\text{E..}}}{3} = 56.4 - \frac{67.9 + 60.5 + 65.1}{3} = -8.1 \]

\[ \text{SE}(\hat{C}) = \sqrt{13.91} \left( \frac{1}{10} + \frac{(1/3)^2}{10} + \frac{(1/3)^2}{10} + \frac{(1/3)^2}{10} \right) \approx 1.362 \]

\[ t-\text{stat} = \frac{\hat{C}_1}{\text{SE}(\hat{C}_1)} = \frac{-8.1}{1.362} \approx -5.948 \]
Scheffé’s critical value for controlling FWER at 0.05 is

\[
\sqrt{(g - 1)F_{\alpha, g-1, \text{df of MSE}}} = \sqrt{(5 - 1)F_{0.05, 5-1, 32}} \\
\approx \sqrt{(5 - 1) \times 2.668} \approx 3.27
\]

\[
> \text{qf}(0.05, \text{df1}=5-1, \text{df2}=32, \text{lower.tail}=F) \\
[1] 2.668437 \\
> \text{sqrt}((5-1)*\text{qf}(0.05, \text{df1}=5-1, \text{df2}=32, \text{lower.tail}=F)) \\
[1] 3.26707
\]

The critical value 3.27 for Scheffé’s method means that: if all treatments (exams) are equal, the contrast of exam effects with the greatest \( t \)-statistic will exceed 3.27 for only 5% of the time. The magnitude of the \( t \)-statistic \(-5.948\) for the contrast we considered is above the critical value 3.72.

Conclusion: We can be certain that the contrast is really significant, even if the contrast was suggested by data snooping.
Tukey’s HSD for Comparing Graders (Problem 13.4)

The SE for comparing the 5 graders pairwise is

\[
SE = \sqrt{\text{MSE} \left( \frac{1}{10} + \frac{1}{10} \right)} = \sqrt{13.91 \times \frac{1}{5}} \approx 1.668.
\]

The critical value for Tukey’s method

\[
> \frac{\text{qtukey}(0.95, 5, 32)}{\sqrt{2}}
\]

[1] 2.889395

Tukey’s HSD = SE × (critical value) \( \approx 1.668 \times 2.889 \approx 4.89 \)

Underline Diagram based on HSD:

\[
\begin{array}{ccccc}
5 & 4 & 3 & 1 & 2 \\
(59.5) & (64.7) & (65.4) & (66.0) & (68.6)
\end{array}
\]

Compared with Underline Diagram based on LSD = 3.40 in HW12

\[
\begin{array}{ccccc}
5 & 4 & 3 & 1 & 2 \\
(59.5) & (64.7) & (65.4) & (66.0) & (68.6)
\end{array}
\]
aov1 = aov(score ~ as.factor(student)+as.factor(grader)+as.factor(exam),data=mydata)

TukeyHSD(aov1)

Tukey multiple comparisons of means
95% family-wise confidence level

$'as.factor(student)'$

<table>
<thead>
<tr>
<th>diff</th>
<th>lwr</th>
<th>upr</th>
<th>p adj</th>
</tr>
</thead>
<tbody>
<tr>
<td>2-1</td>
<td>2.0</td>
<td>-6.009101</td>
<td>10.0091013</td>
</tr>
<tr>
<td>3-1</td>
<td>0.2</td>
<td>-7.809101</td>
<td>8.2091013</td>
</tr>
</tbody>
</table>

...(omitted)...

<table>
<thead>
<tr>
<th>diff</th>
<th>lwr</th>
<th>upr</th>
<th>p adj</th>
</tr>
</thead>
<tbody>
<tr>
<td>10-9</td>
<td>-22.2</td>
<td>-30.209101</td>
<td>-14.1908987</td>
</tr>
</tbody>
</table>

$'as.factor(grader)'$

<table>
<thead>
<tr>
<th>diff</th>
<th>lwr</th>
<th>upr</th>
<th>p adj</th>
</tr>
</thead>
<tbody>
<tr>
<td>2-1</td>
<td>2.6</td>
<td>-2.219533</td>
<td>7.4195329</td>
</tr>
<tr>
<td>3-1</td>
<td>-0.6</td>
<td>-5.419533</td>
<td>4.2195329</td>
</tr>
<tr>
<td>4-1</td>
<td>-1.3</td>
<td>-6.119533</td>
<td>3.5195329</td>
</tr>
<tr>
<td>5-1</td>
<td>-6.5</td>
<td>-11.319533</td>
<td>-1.6804671</td>
</tr>
<tr>
<td>3-2</td>
<td>-3.2</td>
<td>-8.019533</td>
<td>1.6195329</td>
</tr>
<tr>
<td>4-2</td>
<td>-3.9</td>
<td>-8.719533</td>
<td>0.9195329</td>
</tr>
<tr>
<td>5-2</td>
<td>-9.1</td>
<td>-13.919533</td>
<td>-4.2804671</td>
</tr>
<tr>
<td>4-3</td>
<td>-0.7</td>
<td>-5.519533</td>
<td>4.1195329</td>
</tr>
<tr>
<td>5-3</td>
<td>-5.9</td>
<td>-10.719533</td>
<td>-1.0804671</td>
</tr>
<tr>
<td>5-4</td>
<td>-5.2</td>
<td>-10.019533</td>
<td>-0.3804671</td>
</tr>
</tbody>
</table>

$'as.factor(exam)'$

...(omitted)...

Chapter 5 - 43
Tukey-Kramer for BIBD

Recall for BIBD, the estimate of $\alpha_{i_1} - \alpha_{i_2}$ is

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

where $Q_i = y_{i\bullet} - \frac{1}{k} \sum_j I_{ij} y_{\bullet j}$ and $I_{ij} = 1$ if treatment $i$ appears in block $j$, or 0 otherwise.

- $\text{SE}(\hat{\alpha}_{i_1} - \hat{\alpha}_{i_2}) = \sqrt{\text{MSE} \left( \frac{2k}{\lambda g} \right)}$

- $t$-statistic = $\frac{\hat{\alpha}_{i_1} - \hat{\alpha}_{i_2}}{\text{SE}}$ with df = df of MSE

- Tukey-Kramer: reject $H_0: \alpha_{i_1} = \alpha_{i_2}$ if

$$|t| > q_\alpha(g, \text{df of MSE})/\sqrt{2}.$$
Recall Problem 14.3 — Exam Grading

<table>
<thead>
<tr>
<th>Exam</th>
<th>Grader</th>
<th>Score</th>
<th>Exam</th>
<th>Grader</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>60</td>
<td>16</td>
<td>1</td>
<td>61</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>64</td>
<td>17</td>
<td>2</td>
<td>78</td>
</tr>
<tr>
<td>3</td>
<td>11</td>
<td>84</td>
<td>18</td>
<td>3</td>
<td>67</td>
</tr>
<tr>
<td>4</td>
<td>16</td>
<td>72</td>
<td>19</td>
<td>4</td>
<td>84</td>
</tr>
<tr>
<td>5</td>
<td>21</td>
<td>65</td>
<td>20</td>
<td>5</td>
<td>81</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>52</td>
<td>21</td>
<td>6</td>
<td>54</td>
</tr>
<tr>
<td>7</td>
<td>2</td>
<td>57</td>
<td>22</td>
<td>7</td>
<td>51</td>
</tr>
<tr>
<td>8</td>
<td>3</td>
<td>55</td>
<td>23</td>
<td>8</td>
<td>32</td>
</tr>
<tr>
<td>9</td>
<td>4</td>
<td>88</td>
<td>24</td>
<td>9</td>
<td>43</td>
</tr>
<tr>
<td>10</td>
<td>5</td>
<td>65</td>
<td>25</td>
<td>10</td>
<td>58</td>
</tr>
<tr>
<td>11</td>
<td>10</td>
<td>79</td>
<td>26</td>
<td>1</td>
<td>66</td>
</tr>
<tr>
<td>12</td>
<td>2</td>
<td>70</td>
<td>27</td>
<td>2</td>
<td>73</td>
</tr>
<tr>
<td>13</td>
<td>3</td>
<td>48</td>
<td>28</td>
<td>3</td>
<td>58</td>
</tr>
<tr>
<td>14</td>
<td>4</td>
<td>75</td>
<td>29</td>
<td>4</td>
<td>95</td>
</tr>
<tr>
<td>15</td>
<td>5</td>
<td>79</td>
<td>30</td>
<td>5</td>
<td>47</td>
</tr>
</tbody>
</table>

- \( g = 25 \) graders (treatments)
- \( b = 30 \) exams (blocks)
- Each exam was graded by 5 graders (size of block \( k = 5 \))
- Each grader graded 6 exams (number of replicates per treatment \( r = 6 \))
- Every pair of graders graded 1 exam in common (\( \lambda = 1 \))
Problem 14.3 — Exam Grading – Tukey’s HSD

How to identify inconsistent graders?

Recall the SE for pairwise comparisons for the grader effects $\alpha_{i_1} - \alpha_{i_2}$ is

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

with df = (df of MSE) = 96.

By Tukey-Kramer: we reject $H_0$: $\alpha_{i_1} = \alpha_{i_2}$ if

$$|t| > q_{\alpha}(g, \text{df of MSE})/\sqrt{2}.$$
Problem 14.3 — Exam Grading

We have obtained $\hat{\alpha}_1, \hat{\alpha}_2, \ldots, \hat{\alpha}_{24}$ in R on p. 21 of Ch14 Slides.

```r
> 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 FWER = 5%, Tukey’s HSD = 6.38)

3  5  16  6  15  14  8  21  9  1  19  23  24  18  10  13  17  25  12  22  7  20  11  2  4

After Tukey’s adjustment, only Grader #3 and # 4 are significantly inconsistent with most other graders.
Grader #2 and #5 were consistent with all the rest except #3 and #4.

Chapter 5 - 47
Problem 14.3 — Exam Grading

Please note that the R function `TukeyHSD()` doesn’t perform Tukey’s adjustment correctly for BIBD.

Do NOT use `TukeyHSD()` on BIBD.