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.

```
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
```

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

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

Pairwise comparisons using t tests with pooled SD

data:  y and trt

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

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 **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 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

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

- 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 θ may fail to cover θ 5% of the time.

What if we construct multiple 95% confidence intervals

$$\{(L_1, U_1), (L_2, U_2), \dots, (L_k, U_k)\}$$

for several different parameters $\theta_1, \theta_2, \dots, \theta_k$, 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

$$\mathrm{P}(L_i \le \theta_i \le U_i \;\text{ for all } i) > 1 - \alpha.$$

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

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

and reject H_0 when

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

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

$$\begin{aligned} \text{FWER} &= \mathrm{P}(\text{at least one of } H_{01}, \ldots, H_{0k} \text{ is rejected}) \\ &\leq \sum_{i=1}^{k} \underbrace{\mathrm{P}(H_{0i} \text{ is rejected})}_{\text{type I error rate for } \mathrm{H}_{0i}} \end{aligned}$$

If the Type I error rate for each of the k nulls can be controlled at α/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 α/k,

or equivalently if

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

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.

Treatment	1N	1Y	2N	3N	4N	4Y
	97	83	85	64	52	48
	96	87	84	72	56	58
	92	78	78	63	44	49
	95	81	79	74	50	53

```
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

```
> 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 $= 17.97$.

```
> 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

$$SE = \sqrt{\text{MSE}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)} = \sqrt{17.97\left(\frac{1}{4} + \frac{1}{4}\right)} \approx 2.9975$$

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

$$t = \frac{\overline{y}_{i\bullet} - \overline{y}_{j\bullet}}{SE} > t_{N-g,\alpha/2/k}$$

where $k = 15$ since there are $\binom{g}{2} = \binom{6}{2} = 15$ pairs to compare. df $= N - g = 24 - 6 = 18$, $t_{N-g,\alpha/2/k} = t_{18,0.05/2/15} \approx 3.38$.

```
> 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

$$|\overline{y}_{i\bullet} - \overline{y}_{j\bullet}| > \text{SE} \times t_{N-g,\alpha/2/k} \approx 2.9975 \times 3.38 \approx 10.13 = \text{BSD}.$$

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

4N	4Y	3N	2N	1Y	1N
50.50	52.00	68.25	81.50	82.25	95.00

Alternatively, one can compute the *P*-values based on the ordinary pairwise *t*-test

```
> 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.

$$\underline{\text{4N} \quad \text{4Y}} \quad \text{3N} \quad \underline{\text{2N} \quad \text{1Y}} \quad \text{1N}$$

```
> 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

   1N      1Y      2N      3N      4N
1Y 0.00717 -       -       -       -
2N 0.00412 1.00000 -       -       -
3N 7.5e-07 0.00286 0.00496 -       -
4N 2.3e-10 5.5e-08 8.0e-08 0.00020 -
4Y 4.1e-10 1.2e-07 1.7e-07 0.00057 1.00000

P value adjustment method: bonferroni
```

Each Bonferorni *P*-value is the corresponding unadjusted *P*-value multiplied k.
You can just compare the Bonferrnoi *P*-values with α.

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 α.

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

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

  in which $q = \frac{\overline{y}_{\max} - \overline{y}_{\min}}{\sqrt{\text{MSE}/n}}$ 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 α for balanced designs ($n_1 = \ldots = n_g$); approximately at α 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

$$\overline{y}_{i\bullet} - \overline{y}_{k\bullet} \pm \frac{q_\alpha(g, N-g)}{\sqrt{2}} \mathrm{SE}(\overline{y}_{i\bullet} - \overline{y}_{k\bullet})$$

For $\mathrm{H}_0 : \mu_i - \mu_k = 0$ v.s. $\mathrm{H}_a : \mu_i - \mu_k \neq 0$, reject H_0 if

$$|t_0| = \frac{|\overline{y}_{i\bullet} - \overline{y}_{k\bullet}|}{\mathrm{SE}(\overline{y}_{i\bullet} - \overline{y}_{k\bullet})} > \frac{q_\alpha(g, N-g)}{\sqrt{2}}$$

In both the C.I. and the test,

$$\mathrm{SE}(\overline{y}_{i\bullet} - \overline{y}_{k\bullet}) = \sqrt{\mathrm{MSE}\left(\frac{1}{n_i} + \frac{1}{n_k}\right)}.$$

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 $q_\alpha(a, f)$: `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

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

4N	4Y	3N	2N	1Y	1N
50.50	52.00	68.25	81.50	82.25	95.00

Tukey's HSD in R

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

```
> 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
        diff        lwr        upr     p adj
1Y-1N -12.75 -22.276745  -3.223255 0.0054014
2N-1N -13.50 -23.026745  -3.973255 0.0031780
3N-1N -26.75 -36.276745 -17.223255 0.0000007
4N-1N -44.50 -54.026745 -34.973255 0.0000000
4Y-1N -43.00 -52.526745 -33.473255 0.0000000
2N-1Y  -0.75 -10.276745   8.776745 0.9998405
3N-1Y -14.00 -23.526745  -4.473255 0.0022319
4N-1Y -31.75 -41.276745 -22.223255 0.0000000
4Y-1Y -30.25 -39.776745 -20.723255 0.0000001
3N-2N -13.25 -22.776745  -3.723255 0.0037927
4N-2N -31.00 -40.526745 -21.473255 0.0000001
4Y-2N -29.50 -39.026745 -19.973255 0.0000002
4N-3N -17.75 -27.276745  -8.223255 0.0001661
4Y-3N -16.25 -25.776745  -6.723255 0.0004624
```

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 -22.276745  -3.223255 0.0054014
2N-1N -13.50 -23.026745  -3.973255 0.0031780
3N-1N -26.75 -36.276745 -17.223255 0.0000007
4N-1N -44.50 -54.026745 -34.973255 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 ≈ 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

$$\widehat{C} = \sum_{i=1}^{g} \omega_i \overline{y}_{i\bullet}, \quad \mathrm{SE}(\widehat{C}) = \sqrt{\mathrm{MSE} \times \sum_{i=1}^{g} \frac{\omega_i^2}{n_i}}$$

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

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

- For testing $\mathrm{H}_0 : C = 0$ v.s. $\mathrm{H}_a : C \neq 0$, reject H_0 when

$$|t_0| = \frac{|\widehat{C}|}{\mathrm{SE}(\widehat{C})} > \sqrt{(g-1)F_{\alpha,g-1,N-g}}$$

Scheffé's Method for Comparing All Contrasts

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

- Controls (strong) FWER at α,
  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

$$\widehat{C} = \sum_{i=1}^{g} \omega_i \overline{y}_{i\bullet} = \sum_{i=1}^{g} \omega_i (\overline{y}_{i\bullet} - \overline{y}_{\bullet\bullet}).$$

By the Cauchy-Schwartz Inequality $|\sum a_i b_i| \le \sqrt{\sum a_i^2 \sum b_i^2}$ and let $a_i = \dfrac{\omega_i}{\sqrt{n_i}}$ and $b_i = \sqrt{n_i}(\overline{y}_{i\bullet} - \overline{y}_{\bullet\bullet})$, we get

$$|\widehat{C}| = \left| \sum_{i=1}^{g} \omega_i (\overline{y}_{i\bullet} - \overline{y}_{\bullet\bullet}) \right| \le \sqrt{\sum_{i=1}^{g} \frac{\omega_i^2}{n_i} \sum_{i=1}^{g} n_i (\overline{y}_{i\bullet} - \overline{y}_{\bullet\bullet})^2}$$

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

$$|\widehat{C}| \le \sqrt{\sum_{i=1}^{g} \frac{\omega_i^2}{n_i} SS_{trt}}.$$

Proof of Scheffé's Method (2)

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

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

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

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

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

$$|t_0(C)| \le \sqrt{(g-1)F}.$$

Proof of Scheffé's Method (3)

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

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

$$\begin{aligned} FWER &= \mathrm{P}\left(|t_0(C)| > \sqrt{(g-1)F_{\alpha,g-1,N-g}} \text{ for any contrast} C\right) \\ &\leq \mathrm{P}\left(\sqrt{(g-1)F} > \sqrt{(g-1)F_{\alpha,g-1,N-g}}\right) \\ &= \mathrm{P}(F > F_{\alpha,g-1,N-g}) = \alpha. \end{aligned}$$

A Contrast for Nitrogen Effect

Group	1N	1Y	2N	3N	4N	4Y
$\overline{y}_{i\bullet}$	95	82.25	81.5	68.25	50.5	52,

MSE = 17.97

The contrast we consider is

$$C = \frac{\mu_{1\text{N}} + \mu_{1\text{Y}}}{2} - \frac{\mu_{4\text{N}} + \mu_{4\text{Y}}}{2}$$

which is estimated to be

$$\widehat{C} = \frac{\overline{y}_{1N\bullet} + \overline{y}_{1Y\bullet}}{2} - \frac{\overline{y}_{4N\bullet} + \overline{y}_{4Y\bullet}}{2} = \frac{95 + 82.25}{2} - \frac{50.5 + 52}{2} = 37.375.$$

with the standard error

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

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

$$t = \frac{\widehat{C}}{\text{SE}(\widehat{C})} \approx \frac{37.375}{2.12} \approx 17.63.$$

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)

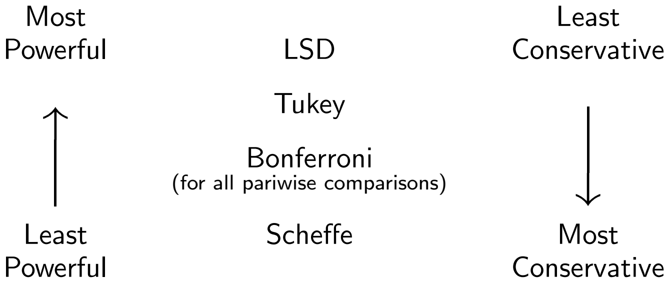
Summary of Multiple Comparison Adjustments

Method	*Family of Tests*	*Critical Value to Keep FWER* $< \alpha$
Fisher's LSD	a single pairwise comparison	$t_{\alpha/2,N-g}$
Tukey-Kramer	all pairwise comparisons	$q_\alpha(g, N-g)/\sqrt{2}$
Bonferroni	varies	$t_{\alpha/(2k),N-g}$, where $k = \#$ of tests
Scheffé	all contrasts	$\sqrt{(g-1)F_{\alpha,g-1,N-g}}$

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

Significance Level vs. Power



Most Powerful

LSD

Least Conservative

Tukey

Bonferroni
(for all pariwise comparisons)

Scheffe

Least Powerful

Most Conservative

In the figure above, Bonferroni is the Bonferroni for all pairwise comparisons.

For a smaller family of, say k tests, one can divide α 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

Method	*Family of Tests*	*Critical Value to Keep FWER* $< \alpha$
Fisher's LSD	a single pairwise comparison	$t_{\alpha/2,\text{df of MSE}}$
Tukey-Kramer	all pairwise comparisons	$q_\alpha(g, \text{df of MSE})/\sqrt{2}$
Bonferroni	all pairwise comparisons	$t_{\alpha/(2r),\text{df of MSE}}$, where $r = \frac{g(g-1)}{2}$
Scheffé	all contrasts	$\sqrt{(g-1)F_{\alpha,g-1,\text{df of MSE}}}$

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

	Plant				
	1	2	3	4	5
Water	-9	18	10	9	-6
	-6	5	9	0	13
Spores	-4	29	4	-2	11
	7	10	-1	6	-1
Oil	4	29	14	14	7
	11	36	16	18	15

Example 13.1 (Mealybugs on Cycads)

Treatment	Water	Spore	Oil
$\bar{y}_{i\bullet}$	4.3	5.9	16.4

MSE = 17.725
df of MSE = 8

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 $2.857 \times 2.663 \approx 7.608$.

$$\underline{\text{Water}\qquad\text{Spore}}\qquad\text{Oil}$$

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)

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

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

$trt
            diff       lwr       upr     p adj
Spore-Water  1.6 -6.008532  9.208532 0.8235730
Oil-Water   12.1  4.491468 19.708532 0.0047478
Oil-Spore   10.5  2.891468 18.108532 0.0105848

$‘as.factor(plant)‘
          diff        lwr        upr     p adj
2-1  20.666667   8.790833 32.5425005 0.0021283
3-1   8.166667  -3.709167 20.0425005 0.2154812
4-1   7.000000  -4.875834 18.8758339 0.3302742
5-1   6.000000  -5.875834 17.8758339 0.4607553
3-2 -12.500000 -24.375834 -0.6241661 0.0390953
4-2 -13.666667 -25.542501 -1.7908328 0.0248443
5-2 -14.666667 -26.542501 -2.7908328 0.0169882
4-3  -1.166667 -13.042501 10.7091672 0.9965298
5-3  -2.166667 -14.042501  9.7091672 0.9657205
5-4  -1.000000 -12.875834 10.8758339 0.9980873
```

- 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`
- `grader`s are reused but `student`s are not
- treatment: `exam`

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

Model: $$\underset{\text{(score)}}{y_{ijk}} = \mu + \underset{\text{(exam)}}{\alpha_i} + \underset{\text{(student)}}{\beta_j} + \underset{\text{(grader)}}{\gamma_k} + \varepsilon_{ijk}$$

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
                   Df Sum Sq Mean Sq F value    Pr(>F)
as.factor(student)  9 5050.3  561.15 40.3376 5.785e-15 ***
as.factor(grader)   4  443.3  110.83  7.9669 0.0001417 ***
as.factor(exam)     4 1889.9  472.48 33.9639 4.246e-11 ***
Residuals          32  445.2   13.91
> library(mosaic)
> mean(score ~ exam, data=mydata)
   1    2    3    4    5
56.4 74.3 67.9 60.5 65.1
```

$$\widehat{C} = \bar{y}_{A\bullet\bullet} - \frac{\bar{y}_{C\bullet\bullet} + \bar{y}_{D\bullet\bullet} + \bar{y}_{E\bullet\bullet}}{3} = 56.4 - \frac{67.9 + 60.5 + 65.1}{3} = -8.1$$

$$\mathrm{SE}(\widehat{C}) = \sqrt{13.91}\sqrt{\frac{1}{10} + \frac{(1/3)^2}{10} + \frac{(1/3)^2}{10} + \frac{(1/3)^2}{10}} \approx 1.362$$

$$t\text{-stat} = \frac{\widehat{C}_1}{\mathrm{SE}(\widehat{C}_1)} = \frac{-8.1}{1.362} \approx -5.948$$

Scheffé's critical value for controlling FWER at 0.05 is

$$\begin{aligned}\sqrt{(g-1)F_{\alpha,g-1,\text{df of MSE}}} &= \sqrt{(5-1)F_{0.05,5-1,32}}\\ &\approx \sqrt{(5-1)\times 2.668} \approx 3.27\end{aligned}$$

```
> qf(0.05, df1=5-1, df2=32, lower.tail=F)
[1] 2.668437
> sqrt((5-1)*qf(0.05, df1=5-1, df2=32, 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

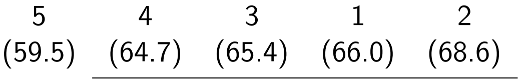
$$\text{SE} = \sqrt{\text{MSE}\left(\frac{1}{10} + \frac{1}{10}\right)} = \sqrt{13.91 \times \frac{1}{5}} \approx 1.668.$$

The critical value for Tukey's method

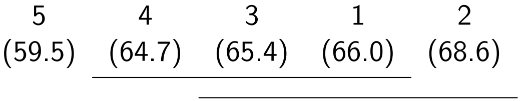
```
> qtukey(0.95, 5, 32)/sqrt(2)
[1] 2.889395
```

Tukey's HSD $=$ SE $\times$ (critical value) $\approx 1.668 \times 2.889 \approx 4.89$

Underline Diagram based on HSD:



Compared ww/ Underline Diagram based on LSD $= 3.40$ in HW12



```
aov1 = aov(score ~ as.factor(student)+as.factor(grader)+as.factor(exam),data=my
TukeyHSD(aov1)

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

$`as.factor(student)`
      diff        lwr         upr     p adj
2-1    2.0  -6.009101  10.0091013 0.9970289
3-1    0.2  -7.809101   8.2091013 1.0000000
...(omitted)...
10-9 -22.2 -30.209101 -14.1908987 0.0000000

$`as.factor(grader)`
    diff        lwr        upr     p adj
2-1  2.6  -2.219533  7.4195329 0.5335412
3-1 -0.6  -5.419533  4.2195329 0.9962326
4-1 -1.3  -6.119533  3.5195329 0.9347241
5-1 -6.5 -11.319533 -1.6804671 0.0039855
3-2 -3.2  -8.019533  1.6195329 0.3285045
4-2 -3.9  -8.719533  0.9195329 0.1593064
5-2 -9.1 -13.919533 -4.2804671 0.0000493
4-3 -0.7  -5.519533  4.1195329 0.9931866
5-3 -5.9 -10.719533 -1.0804671 0.0102914
5-4 -5.2 -10.019533 -0.3804671 0.0293100

$`as.factor(exam)`
...(omitted)...
```

Tukey-Kramer for BIBD

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

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

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.

- $\text{SE}(\widehat{\alpha}_{i_1} - \widehat{\alpha}_{i_2}) = \sqrt{\text{MSE}\left(\frac{2k}{\lambda g}\right)}$
- $t\text{-statistic} = \dfrac{\widehat{\alpha}_{i_1} - \widehat{\alpha}_{i_2}}{\text{SE}}$ with df = df of MSE
- Tukey-Kramer: reject H_0: $\alpha_{i_1} = \alpha_{i_2}$ if

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

```
> qtukey(1-alpha, g, df = df of MSE)/sqrt(2)
```

Recall Problem 14.3 — Exam Grading

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

- $g = 25$ 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

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

```
> qtukey(0.95, 25, df = 96)/sqrt(2)
[1] 3.767619
```

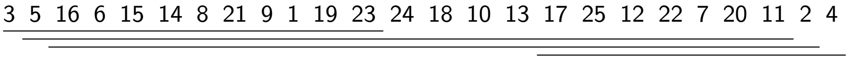
$$\text{Tukey's HSD} = \frac{q_{0.05}(25, 96)}{\sqrt{2}} SE = 3.768 \times 1.6935 \approx 6.38.$$

Problem 14.3 — Exam Grading

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

```
> 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.

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.