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
- Dunnett's Procedure for Comparing with a Control
- Scheffè'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.

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 just do one test.
- If a comparison or contrast is determined after looking at the data (data snooping), one must adjust for multiple comparison.

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

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

Note here that L_i's and U_i's are random variables that depends on the data.

Multiple Comparisons

To account for the fact that we are actually doing multiple comparison, we will need to make our C.I. wider, and the critical value 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 the same as in the usual t-tests and t-intervals. Only the critical values vary with methods, as summarized on Slide 19.

5.2 Bonferroni's Method

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

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

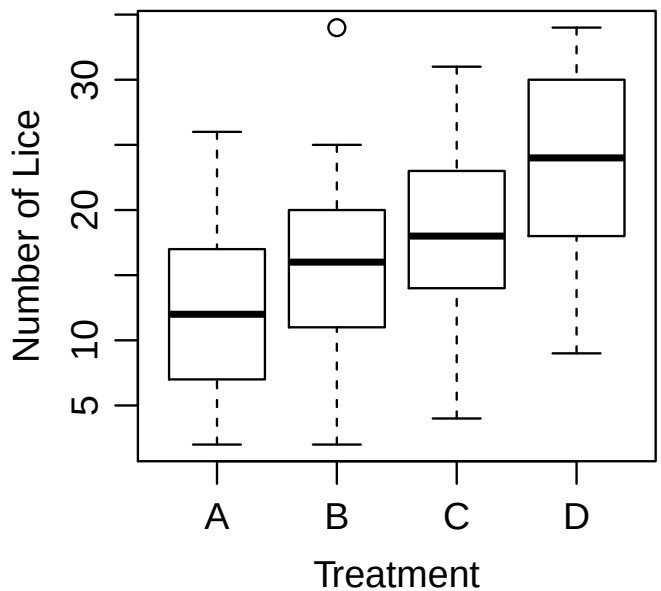
If the Type I error rate for each of the k nulls can be controlled at α/k, then

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

- Bonferroni's method rejects a null if the comparisonwise P-value is less than α/k
- Bonferroni's method works OK when k is small
- When $k > 10$, Bonferroni starts to get too conservative than necessary.
  The actual FWER can be much less than α.

Example: Beet Lice Study

- Goal: efficacy of 4 chemical treatments for beet lice
- 100 beet plants in individual pots in total, 25 plants per treatment, randomly assigned
- Response: # of lice on each plant at the end of the 2nd week
- The pots are spatially separated
- Data file: `beetlice.txt` is posted on Canvas with the slides



Example — Beet Lice

The group means of the 4 treatments are

```
> beet = read.table("beetlice.txt", header=TRUE)
> library(mosaic)
> mean(licecount ~ ttt, data = beet)
    A     B     C     D
12.00 14.96 18.36 24.00
```

From the ANOVA table below, we get MSE $= 47.8$.

```
> lm1 = lm(licecount ~ ttt, data = beet)
> anova(lm1)
           Df Sum Sq Mean Sq F value   Pr(>F)
ttt         3   1989   663.1   13.86 1.39e-07 ***
Residuals  96   4593    47.8
```

The SE for pairwise comparison is

$$SE = \sqrt{\text{MSE}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)} = \sqrt{47.8\left(\frac{1}{25} + \frac{1}{25}\right)} = 1.956$$

Example — Beet Lice (Bonferoni's Method)

Chemical Comparison	Estimate	SE	t-value	p-value of t-test	
$\mu_B - \mu_A$	2.96	1.956	1.513	0.13356	
$\mu_C - \mu_A$	6.36	1.956	3.251	0.00159	< 0.0083
$\mu_D - \mu_A$	12.00	1.956	6.134	1.91×10^{-8}	< 0.0083
$\mu_C - \mu_B$	3.40	1.956	1.738	0.0854	
$\mu_D - \mu_B$	9.04	1.956	4.621	1.19×10^{-5}	< 0.0083
$\mu_D - \mu_C$	5.64	1.956	2.883	0.00486	< 0.0083

There are $k = 6$ tests.

For $\alpha = 0.05$, instead of rejecting a null when the P-value $< \alpha$, Bonferoni's method rejects when

$$\text{the } P\text{-value } < \frac{\alpha}{k} = \frac{0.05}{6} = 0.0083.$$

Only AC, AD, BD, CD are significantly different.

A B C D

Example — Beet Lice (Bonferoni's Method)

Alternatively, 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 $k = 6$ since there are $\binom{g}{2} = \binom{4}{2} = 6$ pairs to compare.
df $= N - g = 100 - 4 = 96$, $t_{N-g,\alpha/2/k} = t_{96,0.05/2/6} \approx 2.694$.

```
> qt(0.05/2/6, df=96, lower.tail=F)
[1] 2.694028
```

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

$$|\overline{y}_{i\bullet} - \overline{y}_{j\bullet}| > \text{SE} \times t_{N-g,\alpha/2/k} \approx 1.956 \times 2.694 \approx 5.27 = \text{BSD}.$$

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

A	B	C	D
12.00	14.96	18.36	24.00

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

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

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

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

In both the C.I. and the test,

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

Tukey's HSD

To be significant at FWER $= \alpha$ based Tukey's correction, the mean difference between a pair of treatments i and k must be at least

$$\frac{q_\alpha(g, N-g)}{\sqrt{2}} \times \sqrt{\text{MSE}\left(\frac{1}{n_i} + \frac{1}{n_k}\right)}$$

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

R command to find $q_\alpha(a, f)$: `qtukey(1-alpha,a,f)`

```
> qtukey(0.95, 4, 96)/sqrt(2)
[1] 2.614607
```

For the Beet Lice example, Tukey's HSD is $2.6146 \times 1.956 \approx 5.114$

Chemical	A	B	C	D
$\overline{y}_{i\bullet}$	12.00	14.96	18.36	24.00

A B C D

Tukey's HSD in R

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

```
> aov1 = aov(licecount ~ ttt, data = beet)
> TukeyHSD(aov1)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = licecount ~ ttt, data = beet)

$ttt
     diff          lwr      upr     p adj
B-C -3.40  -8.5150601  1.71506 0.3099554
A-C -6.36 -11.4750601 -1.24494 0.0084911
D-C  5.64   0.5249399 10.75506 0.0246810
A-B -2.96  -8.0750601  2.15506 0.4337634
D-B  9.04   3.9249399 14.15506 0.0000695
D-A 12.00   6.8849399 17.11506 0.0000001
```

Note that the widths of all CIs above are 2x of the HSD.

E.g., the width of the CI for B-C is $1.71506 - (-8.5150601) = 10.23012$ is twice of HSD $= 5.114$.

5.5.1 Dunnett's Procedure for Comparing with a Control

- Family: comparing ALL TREATMENTS with a CONTROL, $\mu_i - \mu_{\text{ctrl}}$, where μ_{ctrl} is the mean of the control group
- Controls the (strong) FWER *exactly* at α for balanced designs ($n_1 = \ldots = n_g$); approximately at α for unbalanced designs
- Less conservative and greater power than Tukey-Kramer's
- $100(1-\alpha)\%$ Dunnett's simultaneous C.I. for $\mu_i - \mu_{\text{ctrl}}$ is

$$\overline{y}_{i\bullet} - \overline{y}_{\text{control}\bullet} \pm d_\alpha(g-1, N-g)\sqrt{\text{MSE} \times \left(\frac{1}{n_i} + \frac{1}{n_{\text{ctrl}}}\right)}$$

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

$$|t_0| = \frac{|\overline{y}_{i\bullet} - \overline{y}_{a\bullet}|}{\sqrt{\text{MSE} \times \left(\frac{1}{n_i} + \frac{1}{n_a}\right)}} > d_\alpha(g-1, N-g)$$

- The critical values $d_\alpha(g-1, N-g)$ can be found in Table D.9, p.635-638, of the textbook

5.3 Scheffè'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

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

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

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

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

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

Scheffè'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 Scheffè'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 $|\sum a_i b_i| \leq \sqrt{\sum a_i^2 \sum b_i^2}$ 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

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

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

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

Proof of Scheffè's Method (2)

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

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

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

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

Proof of Scheffè'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

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

Example — Beet Lice

Chemical	A	B	C	D
$\overline{y}_{i\bullet}$	12	14.96	18.36	24

MSE = 47.8

Consider a contrast comparing treatment A, B, C (all liquid) with treatment D (powder): $C = \frac{\mu_A + \mu_B + \mu_C}{3} - \mu_D$.

which is estimated by

$$\widehat{C} = \frac{\overline{y}_{A\bullet} + \overline{y}_{B\bullet} + \overline{y}_{C\bullet}}{3} - \overline{y}_{D\bullet} = \frac{12 + 14.96 + 18.36}{3} - 24 = -8.893$$

with standard error

$$\text{SE} = \sqrt{\text{MSE}\sum_{i=1}^{g}\frac{\omega_i^2}{n_i}} = \sqrt{47.8(\frac{(1/3)^2}{25} + \frac{(1/3)^2}{25} + \frac{(1/3)^2}{25} + \frac{(-1)^2}{25})}$$

$$= \sqrt{47.8 \times \frac{4}{75}} = 1.597.$$

t-statistic: $t_0 = \frac{\widehat{C}}{\text{SE}} = \frac{-8.893}{1.597} = -5.568$, with df $= 100-4 = 96$.

Example — Beet Lice

With Scheffè's Method, the critical value controlling FWER at 0.05 is

$$\sqrt{(g-1)F_{\alpha,g-1,N-g}} = \sqrt{(4-1)F_{0.05,3,96}} = \sqrt{(4-1)\times 2.699} \approx 2.846$$

```
> qf(0.05, df1=3, df2=96, lower.tail=F)
[1] 2.699393
> sqrt((4-1)*qf(0.05, df1=3, df2=96, lower.tail=F))
[1] 2.84573
```

The critical value 2.846 for Scheffè's method means that: if all treatments are equal, the contrast with the greatest t-statistic will exceed 2.846 for only 5% of the time. The magnitude of the t-statistic -5.568 for the contrast we considered is far above the critical value 2.846.

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	$t_{\alpha/2,N-g}$
Dunnett	all comparisons with a control	$d_\alpha(g-1, N-g)$
Tukey-Kramer	all pairwise comparisons	$q_\alpha(g, N-g)/\sqrt{2}$
Bonferroni	varies	$t_{\alpha/(2k),N-g}$, where $k = \#$ of tests
Scheffè	all contrasts	$\sqrt{(g-1)F_{\alpha,g-1,N-g}}$

Example — Beet Lice

Recall

treatment	A	B	C	D
n_i	25	25	25	25
$\overline{y}_{i\bullet}$	12.00	14.96	18.36	24.00

, $\text{MSE} = 47.84$.

$$\text{SE}(\overline{y}_{i\bullet} - \overline{y}_{k\bullet}) = \sqrt{\text{MSE}\left(\frac{1}{n_i} + \frac{1}{n_k}\right)} = \sqrt{47.84 \times \frac{2}{25}} = 1.9563.$$

The critical values at $\alpha = 0.05$ are

```
> alpha = 0.05
> g = 4
> r = g*(g-1)/2
> N = 100
> qt(1-alpha/2, df = N-g)                          # Fisher's LSD
[1] 1.984984
> qt(1-alpha/2/r, df = N-g)                        # Bonferroni
[1] 2.694028
> qtukey(1-alpha, g, df = N-g)/sqrt(2)             # Tukey's HSD
[1] 2.614607
> sqrt((g-1)*qf(1-alpha, df1=g-1, df2=N-g)) # Scheffe
[1] 2.84573
```

The half widths of the C.I. are "critical values" $\times$SE, which are

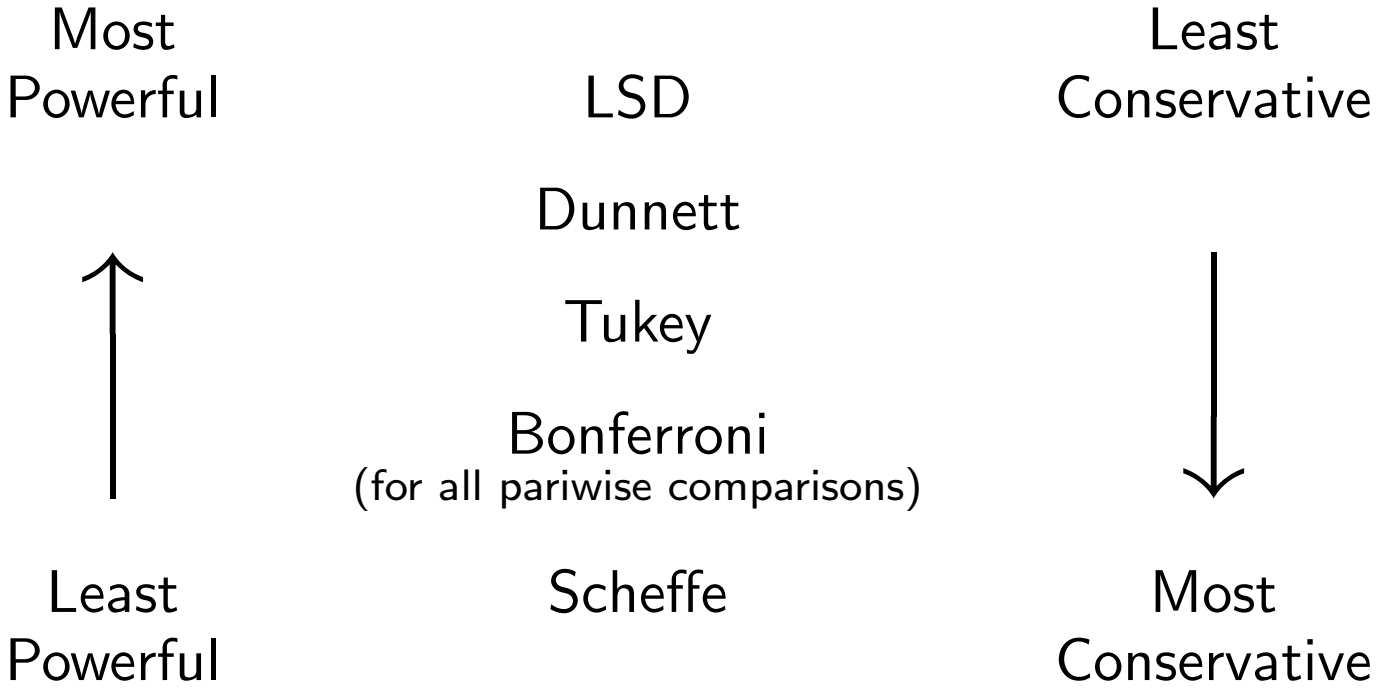
Procedure	LSD	Tukey	Bonferroni	Scheffe
C.I. half width	3.883	5.115	5.270	5.567

	diff	LSD	Tukey	Bonferroni	Scheffe
B-C	-3.40	(-7.28, 0.48)	(-8.51, 1.71)	(-8.67, 1.87)	(-8.97, 2.17)
A-C	-6.36	(-10.24, -2.48)	(-11.47, -1.25)	(-11.63, -1.09)	(-11.93, -0.79)
D-C	5.64	(1.76, 9.52)	(0.53,10.75)	(0.37,10.91)	(0.07,11.21)
A-B	-2.96	(-6.84, 0.92)	(-8.07, 2.15)	(-8.23, 2.31)	(-8.53, 2.61)
D-B	9.04	(5.16,12.92)	(3.93,14.15)	(3.77,14.31)	(3.47,14.61)
D-A	12.00	(8.12,15.88)	(6.89,17.11)	(6.73,17.27)	(6.43,17.57)

Which Procedures to Use?

- Use BONFERRONI when only interested in a small number of planned contrasts (or pairwise comparisons)
- Use DUNNETT when only interested in comparing all treatments with a control
- 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



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	$t_{\alpha/2,\text{df of MSE}}$
Dunnett	all comparisons with a control	$d_\alpha(g-1, \text{df of MSE})$
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 $r = \frac{g(g-1)}{2}$
Scheffè	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 $\times$ 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
$\overline{y}_{i\bullet}$	4.3	5.9	16.4

$\text{MSE} = 17.725$
df of $\text{MSE} = 8$

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 $2.857 \times 2.663 \approx 7.608$.

Water	Spore	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.

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 $Q_i = y_{i\bullet} - \frac{1}{k}\sum_j l_{ij} y_{\bullet j}$ and $l_{ij} = 1$ 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-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

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

Recall Problem 14.3 — Exam Grading

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

- $g = 25$ 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.