STAT 222 Lecture 5-6
Multiple Comparisons

Yibi Huang

Outline

Textbook Coverage: Section 4.4

- Why Worry About Multiple Comparisons?
- Familywise Error Rate (FEWR)
- Simultaneous Confidence Intervals
- 4.4.2 Bonferroni's Method for Pre-planned Comparisons
- 4.4.3 Scheffe's Method for Comparing All Contrasts
- 4.4.4 Tukey's Method for Pairwise Comparisons

Skip 4.4.5 Dunnett Method for Treatment-Versus-Control Comparisons

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

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?

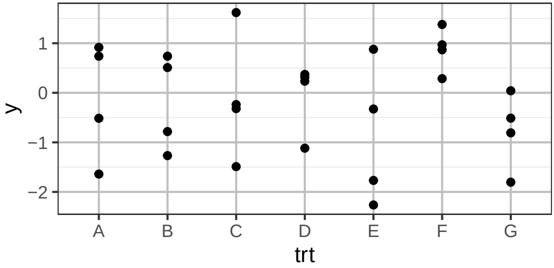
- When 100 H_0's are tested at 0.05 level, even if all H_0's are true, it's normal to have 5 being rejected.
- When multiple tests are done, it's very likely that some significant results may be NOT be TRUE FINDINGS. The significance must be adjusted

Why Worry About Multiple Comparisons?

- In an experiment, when the ANOVA F-test is rejected, we'd then 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 for about 5% of the time.

7 groups of observations of size 4 each are generated from the $N(0, 1)$ distribution.

```
g = 7      # number of treatments
n = 4      # number of replicates per treatment
trt = gl(g, n, labels=LETTERS[1:g])  # Treatment: A, B, C, D, E, F, G
y = rnorm(g*n, mean=0, sd = 1)       # Standard normal
dat.tmp = data.frame(trt,y)
library(ggplot2)
ggplot(dat.tmp, aes(x=trt, y=y)) + geom_point()
```



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

```
library(emmeans)
mod1 = aov(y ~ trt)
mod1emm = emmeans(mod1, "trt")
pwpm(mod1emm, adjust="none")
          A         B         C         D         E         F         G
A [-0.1251]    0.9181    0.9805    0.9177    0.3162    0.1828    0.3836
B    0.0755 [-0.2006]    0.8987    0.8367    0.3667    0.1532    0.4408
C   -0.0180   -0.0935 [-0.1071]    0.9371    0.3050    0.1905    0.3708
D   -0.0759   -0.1515   -0.0579 [-0.0492]    0.2707    0.2169    0.3313
E    0.7452    0.6696    0.7632    0.8211 [-0.8703]    0.0255    0.8924
F   -1.0000   -1.0755   -0.9820   -0.9240   -1.7451 [ 0.8749]    0.0340
G    0.6458    0.5703    0.6638    0.7218   -0.0993    1.6458 [-0.7709]

Row and column labels: trt
Upper triangle: P values
Diagonal: [Estimates] (emmean)
Lower triangle: Comparisons (estimate)   earlier vs. later
```

Repeat the following several times.

```
g = 7   # number of treatments
n = 4    # number of replicates per treatment
trt = gl(g, n, labels=LETTERS[1:g])  # Treatment: A, B, C, D, E, F, G
y = rnorm(g*n, mean=0, sd = 1)       # Standard normal
mod1 = aov(y ~ trt)
mod1emm = emmeans(mod1, "trt")
pwpm(mod1emm, adjust="none")
```

How often do you see a significant difference?

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 **significance level** or **Type I error rate** of a test is the chance of making a Type I error.

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 **significance level** 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

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 **significance level** 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*, or *overall significance level* is defined as the chance of making a familywise Type I error

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

- FWER depends on the family.
  The larger the family, the larger the FWER.

Simultaneous Confidence Intervals

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

Simultaneous Confidence Intervals

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

When 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 can be 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 multiple comparisons, we need to make our C.I.'s 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.

What is Data Snooping

If one looks at data first and decides which contrast(s) to test based on what they see, that is *data snooping*, e.g.,

- when one decides to compare treatment A & E since A has the highest mean and E the lowest
- one decides to test the contrast

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

since A and C have higher means than B and D

What is Data Snooping

If one looks at data first and decides which contrast(s) to test based on what they see, that is *data snooping*, e.g.,

- when one decides to compare treatment A & E since A has the highest mean and E the lowest
- one decides to test the contrast

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

  since A and C have higher means than B and D

Why is Data Snooping Problematic?

- 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.

4.4.2 Bonferroni's Method for Pre-planned Comparisons

Bonferroni's Simultaneous CIs for Pre-planned Contrasts

Consider a multi-sample problem

$$y_{ij} = \mu_i + \varepsilon_{ij}, \quad i = 1, \ldots, g, \quad j = 1, \ldots, n_i.$$

Given m *pre-planned* contrasts: $C_1, C_2, \ldots, C_m$, where $C_k = \sum_{i=1}^{g} c_{ik}\mu_i$ for $k = 1, \ldots, m$, recall the estimates and SE's of these contrasts are respectively.

$$\widehat{C}_k = \sum_{i=1}^{g} c_{ik}\bar{y}_{i\bullet} \quad \text{and} \quad \mathrm{SE}(\widehat{C}_k) = \sqrt{\mathsf{MSE} \times \sum_{i=1}^{g} \frac{c_{ik}^2}{n_i}}$$

The $100(1-\alpha)\%$ simultaneous confidence interval for C_k is

$$\widehat{C}_k \pm t_{N-g,\alpha/2/m} \times \mathrm{SE}(\widehat{C}_k) \quad \text{for } k = 1, \ldots, m.$$

Bonferroni's Tests for Multiple Pre-planned Contrasts

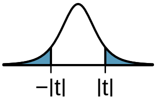
Consider the family of tests for m *pre-planned* contrasts $C_1, \ldots, C_m$,

$$\begin{aligned} H_{01} &: C_1 = 0 \quad \text{v.s.} \quad H_{a1}: \ C_1 \neq 0 \\ &\vdots \\ H_{0m} &: C_m = 0 \quad \text{v.s.} \quad H_{am}: \ C_m \neq 0 \end{aligned}$$

Bonferroni's method rejects H_{0k}: $C_k = 0$ against H_{ak}: $C_k \neq 0$ controlling FWER at α if

$$|t\text{-stat}| = \frac{|\widehat{C}_k|}{\mathrm{SE}(\widehat{C}_k)} > t_{N-g,\alpha/2/m}.$$

or equivalently if

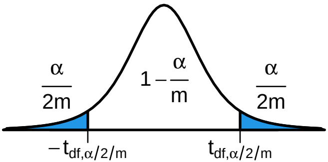
P-value = 

=`2*pt(abs(t), df, lower.tail=F)`$< \dfrac{\alpha}{m}$.

Proof of Bonferroni's Method (1)

If H_{0k}: $C_k = 0$ is true, the t-statistic $= \widehat{C}_k/\mathrm{SE}(\widehat{C}_k)$ has a t-distribution with $N - g$ degrees of freedom, we know

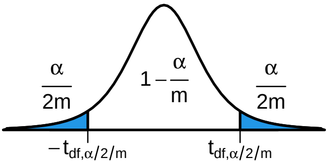
$$\mathrm{P}\left(|t\text{-stat}| > t_{N-g,\alpha/2/m}\right) = \frac{\alpha}{m}$$



Proof of Bonferroni's Method (1)

If H_{0k}: $C_k = 0$ is true, the t-statistic $= \widehat{C}_k/\mathrm{SE}(\widehat{C}_k)$ has a t-distribution with $N - g$ degrees of freedom, we know

$$\mathrm{P}\left(|t\text{-stat}| > t_{N-g,\alpha/2/m}\right) = \frac{\alpha}{m}$$



If we reject H_{0k}: $C_k = 0$ only when $|t\text{-stat}| > t_{N-g,\alpha/2/m}$, the chance we reject H_{0k} when it is true is

$$\mathrm{P}(H_{0k} \text{ is rejected}) = \mathrm{P}\left(|t\text{-stat}| > t_{N-g,\alpha/2/m}\right) = \frac{\alpha}{m}.$$

Proof of Bonferroni's Method (2)

Using Bonferroni's inequality in probability theory,

$$\mathrm{P}(A_1 \text{ or } A_2 \text{ or } \cdots \text{ or } A_m) \leq \mathrm{P}(A_1) + \mathrm{P}(A_2) + \cdots + \mathrm{P}(A_m)$$

given that $H_{01}, \ldots, H_{0m}$ being all true, we have

$$\begin{aligned}\text{FWER} &= \mathrm{P}(\text{at least one of } H_{01}, \ldots, H_{0m} \text{ is rejected}) \\ &\leq \sum_{k=1}^{m} \mathrm{P}(H_{0k} \text{ is rejected}) \\ &= \underbrace{\frac{\alpha}{m} + \cdots + \frac{\alpha}{m}}_{m \text{ times}} = \alpha\end{aligned}$$

Proof of Bonferroni's Method (3)

Likewise, the interval

$$\widehat{C}_k \pm t_{N-g,\alpha/2/m} \times \mathrm{SE}(\widehat{C}_k) \quad \text{for } k = 1, \ldots, m.$$

would only fail to include C_k with probability α/m.

Using Bonferroni's inequality,

$$\begin{aligned} &\mathrm{P}(\text{any of the } m \text{ CI's } \widehat{C}_k \pm t_{N-g,\alpha/2/m}\mathrm{SE}(\widehat{C}_k) \text{ fail to include } C_k) \\ = &\sum_{k=1}^{m} \mathrm{P}(\widehat{C}_k \pm t_{N-g,\alpha/2/m} \times \mathrm{SE}(\widehat{C}_k) \text{ fail to include } C_k) \\ &= \underbrace{\frac{\alpha}{m} + \cdots + \frac{\alpha}{m}}_{m \text{ times}} = \alpha \end{aligned}$$

Example: Grass/Weed Competition

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
y_{ij}	97	83	85	64	52	48
	96	87	84	72	56	58
	92	78	78	63	44	49
	95	81	79	74	50	53
Mean $\bar{y}_{i\bullet}$	95	82.25	81.5	68.25	50.5	52
Size n_i	4	4	4	4	4	4

, MSE $= 17.97$

Example — Grass/Weed (Bonferroni's Method)

Suppose we are ONLY interested in the 3 contrasts below:

$$C_1 = \mu_{1N} - \mu_{1Y}, \quad C_2 = \mu_{4N} - \mu_{4Y}, \quad C_3 = \frac{\mu_{1N} + \mu_{4N}}{2} - \frac{\mu_{1Y} + \mu_{4Y}}{2}.$$

Their estimates and SE's are

	estimate	SE
C_1	$95 - 82.25 = 12.75$	$\sqrt{17.97(\frac{1}{4} + \frac{1}{4})} \approx 2.9977$
C_2	$50.5 - 52 = -1.5$	$\sqrt{17.97(\frac{1}{4} + \frac{1}{4})} \approx 2.9977$
C_3	$\frac{95+50.5}{2} - \frac{82.25+52}{2} = 5.625$	$\sqrt{17.97(\frac{0.5^2}{4} + \frac{(-0.5)^2}{4} + \frac{0.5^2}{4} + \frac{(-0.5)^2}{4})} \approx 2.11967$

As there are $m = 3$ contrasts, the critical value is

```
qt(0.05/2/3, df=18, lower.tail=F)
[1] 2.639
```

Bonferroni's 95% Simultaneous CI for the 3 contrasts are

$$\begin{aligned}
C_1 &: \quad 12.75 \pm 2.639 \times 2.9977 \approx 12.75 \pm 7.91 = (4.84, 20.66) \\
C_2 &: \quad -1.5 \pm 2.639 \times 2.9977 \approx -1.5 \pm 7.91 = (-9.41, 6.41) \\
C_3 &: \quad 5.625 \pm 2.639 \times 2.11967 \approx 5.625 \pm 5.594 = (0.031, 11.219)
\end{aligned}$$

Bonferroni's Method in R

```
grass = read.table(
        "http://www.stat.uchicago.edu/~yibi/s222/grassweed.txt", h=T)
grass$trt = as.factor(grass$trt)
levels(grass$trt)
[1] "1N" "1Y" "2N" "3N" "4N" "4Y"
mod1 = aov(percent ~ trt, data=grass)
mod1emm = emmeans(mod1, "trt")
contrast(mod1emm, list(C1=c(1, -1, 0, 0, 0, 0),
                       C2=c(0, 0, 0, 0, 1, -1),
                       C3=c(1, -1, 0, 0, 1, -1)/2),
         infer=c(T,T), level=0.95, adjust="bonferroni")
 contrast estimate   SE df lower.CL upper.CL t.ratio p.value
 C1          12.75 3.00 18    4.839    20.66   4.253  0.0014
 C2          -1.50 3.00 18   -9.411     6.41  -0.500  1.0000
 C3           5.62 2.12 18    0.031    11.22   2.654  0.0485

Confidence level used: 0.95
Conf-level adjustment: bonferroni method for 3 estimates
P value adjustment: bonferroni method for 3 tests
```

Compared w/ the CIs and the P-values without Bonferroni's adjustment.

```
contrast(mod1emm, list(C1=c(1, -1, 0, 0, 0, 0),
                       C2=c(0, 0, 0, 0, 1, -1),
                       C3=c(1, -1, 0, 0, 1, -1)/2),
         infer=c(T,T), level=0.95, adjust="none")
 contrast estimate   SE df lower.CL upper.CL t.ratio p.value
 C1          12.75 3.00 18     6.45     19.0   4.253  0.0005
 C2          -1.50 3.00 18    -7.80      4.8  -0.500  0.6229
 C3           5.62 2.12 18     1.17     10.1   2.654  0.0162

Confidence level used: 0.95
```

Note Bonferroni's P-value is simply the unadjusted P-value multiplied by m, cap at 1.

Limitation of Bonferroni's Method

- The number of tests *m* must be finite.
- Bonferroni's method works OK when the number of tests *m* is small
- When the number of tests *m* 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.

4.4.3 Scheffe's Method for Comparing All Contrasts

Scheffe's Method for Comparing All Contrasts

For multi-sample data with a total of N observations divided into g groups. Consider a contrast $C = \sum_{i=1}^{g} c_i \mu_i$. Recall

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

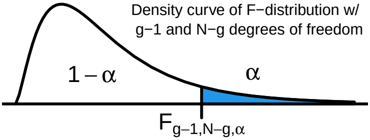
- $100(1-\alpha)\%$ Scheffe's simultaneous C.I. for a contrast C is

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

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

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

where $F_{g-1,N-g,\alpha}$ is the value such that



Scheffe's Method for Comparing All Contrasts

- Controls FWER at α,
  where the family is ALL POSSIBLE CONTRASTS
  $C = \sum_{i=1}^{g} c_i \mu_i$ of the g group means μ_i

- Should be used if one has contrast(s) not pre-planned in advance.
- Protects against data snooping!

Proof of Scheffe's Method (1)

Observe that

$$\widehat{C} = \sum_{i=1}^{g} c_i \bar{y}_{i\bullet} = \left(\sum_{i=1}^{g} c_i \bar{y}_{i\bullet}\right) - \bar{y}_{\bullet\bullet} \underbrace{\sum_{i=1}^{g} c_i}_{=0} = \sum_{i=1}^{g} c_i (\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet}).$$

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

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

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

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

Proof of Scheffe'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 $|\widehat{C}| \le \sqrt{\sum_{i=1}^{g} \frac{c_i^2}{n_i} SS_{trt}}$ proved in the previous page, we have

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

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 Scheffe'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 $|\widehat{C}| \leq \sqrt{\sum_{i=1}^{g} \frac{c_i^2}{n_i} SS_{trt}}$ proved in the previous page, we have

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

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

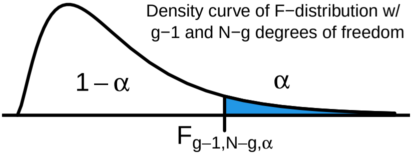
$$|t_0(C)| \leq \sqrt{\frac{SS_{trt}}{MSE}} = \sqrt{\frac{(g-1)MS_{trt}}{MSE}} = \sqrt{(g-1)F}.$$

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 Scheffe's Method (3)

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



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

$$\begin{aligned}
FWER &= \mathrm{P}(\mathrm{H}_0\text{: } C = 0 \text{ is rejected for any contrast } C) \\
&= \mathrm{P}\left(|t_0(C)| > \sqrt{(g-1)F_{\alpha,g-1,N-g}} \text{ for any contrast } C\right) \\
&\le \mathrm{P}\left(\sqrt{(g-1)F} > \sqrt{(g-1)F_{\alpha,g-1,N-g}}\right) \\
&= \mathrm{P}(F > F_{\alpha,g-1,N-g}) = \alpha.
\end{aligned}$$

A Contrast for Nitrogen Effect

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

, MSE = 17.97

The contrast we consider is

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

in which $(c_{1\mathrm{N}}, c_{1\mathrm{Y}}, c_{2\mathrm{N}}, c_{3\mathrm{N}}, c_{4N}, c_{4\mathrm{Y}}) = (0.5, 0.5, 0, 0, -0.5, -0.5)$, estimated to be

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

A Contrast for Nitrogen Effect

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

, MSE $= 17.97$

The contrast we consider is

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

in which $(c_{1\mathrm{N}}, c_{1\mathrm{Y}}, c_{2\mathrm{N}}, c_{3\mathrm{N}}, c_{4N}, c_{4\mathrm{Y}}) = (0.5, 0.5, 0, 0, -0.5, -0.5)$, estimated to be

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

with the standard error

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

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

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

A Contrast for Nitrogen Effect

Scheffe's critical value controlling FWER at 0.05 is

$$\sqrt{(g-1)F_{g-1,N-g,\alpha}} = \sqrt{(6-1)F_{6-1,24-6,0.05}}$$
$$\approx \sqrt{(6-1)\times 2.773} \approx 3.723$$

```
qf(0.05, df1=6-1, df2=24-6, lower.tail=F)
[1] 2.773
sqrt((6-1)*qf(0.05, df1=6-1, df2=24-6, lower.tail=F))
[1] 3.723
```

- Scheffe's critical value 3.723 means that: if all treatments are equal, the contrast with the greatest t-statistic will exceed 3.723 for only 5% of the time
- The t-statistic 17.63 for the nitrogen effect contrast is far above the critical value 3.723
- Conclusion: The contrast is highly significant, even if the contrast was suggested by data snooping

Scheffe's Simultaneous CI for Contrasts

If the Nitrogen effect contrast

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

is considered after looking at the data (suggested by data snooping), it's safer to use Scheffe's 95% simultaneous CI

$$\begin{aligned}\widehat{C} \pm \sqrt{(g-1)F_{g-1,N-g,\alpha}}\,\mathrm{SE}(\widehat{C}) &\approx 37.375 \pm 3.723 \times 2.12 \\ &\approx 37.375 \pm 7.893 \approx (29.482, 45.268)\end{aligned}$$

where the critical value is $\sqrt{(g-1)F_{g-1,N-g,\alpha}} = \sqrt{(6-1)F_{6-1,24-6,0.05}} \approx \sqrt{(6-1) \times 2.773} \approx 3.723.$

```
sqrt((6-1)*qf(0.05, df1=6-1, df2=24-6, lower.tail=F))
[1] 3.723
```

Scheffe's Method in R

To get the correct Scheffe's simultaneous CI and P-value, one must

- specify `adjust ="scheffe"`,
- apply `summary()` on `contrast()`, and
- specify `scheffe.rank = g-1`

```
summary(contrast(mod1emm, list(C4=c(1, 1, 0, 0, -1, -1)/2),
         infer=c(T,T), level=0.95, adjust="scheffe"), scheffe.rank = 5)
 contrast estimate   SE df lower.CL upper.CL t.ratio p.value
 C4           37.4 2.12 18     29.5     45.3  17.632  <.0001

Confidence level used: 0.95
Conf-level adjustment: scheffe method with rank 5
P value adjustment: scheffe method with rank 5
```

Without doing all three things above, the reported simultaneous CI and P-value are not correct, as shown in the next page.

Incorrect Scheffe's Method in R

Not applying `summary()` on `contrast()`, and not specifying `scheffe.rank = g-1`:

```
contrast(mod1emm, list(C4=c(1, 1, 0, 0, -1, -1)/2),
         infer=c(T,T), level=0.95, adjust="scheffe")
 contrast estimate   SE df lower.CL upper.CL t.ratio p.value
 C4           37.4 2.12 18     32.9     41.8  17.632  <.0001

Confidence level used: 0.95
Conf-level adjustment: scheffe method with rank 1
P value adjustment: scheffe method with rank 1
```

Not applying `summary()` on `contrast()`

```
contrast(mod1emm, list(C4=c(1, 1, 0, 0, -1, -1)/2),
         infer=c(T,T), level=0.95, adjust="scheffe", scheffe.rank = 5)
 contrast estimate   SE df lower.CL upper.CL t.ratio p.value
 C4           37.4 2.12 18     32.9     41.8  17.632  <.0001

Confidence level used: 0.95
Conf-level adjustment: scheffe method with rank 1
P value adjustment: scheffe method with rank 1
```

Scheffe's Method in R on Several Contrasts

Recall we considered the 3 contrasts below when demonstrating the Bonferroni's method.

$$C_1 = \mu_{1N} - \mu_{1Y}, \quad C_2 = \mu_{4N} - \mu_{4Y}, \quad C_3 = \frac{\mu_{1N} + \mu_{4N}}{2} - \frac{\mu_{1Y} + \mu_{4Y}}{2}.$$

```
summary(contrast(mod1emm, list(C1=c(1, -1, 0, 0, 0, 0),
                               C2=c(0, 0, 0, 0, 1, -1),
                               C3=c(1, -1, 0, 0, 1, -1)/2),
         infer=c(T,T), level=0.95, adjust="scheffe"), scheffe.rank=5)
 contrast estimate   SE df lower.CL upper.CL t.ratio p.value
 C1          12.75 3.00 18     1.59    23.91   4.253  0.0193
 C2          -1.50 3.00 18   -12.66     9.66  -0.500  0.9982
 C3           5.62 2.12 18    -2.27    13.52   2.654  0.2683

Confidence level used: 0.95
Conf-level adjustment: scheffe method with rank 5
P value adjustment: scheffe method with rank 5
```

Note C_3 is not significant after Scheffe's adjustment though it's significant after Bonferroni's adjustment of 3 contrasts

4.4.4 Tukey Method for Pairwise Comparisons

4.4.4 Tukey Method for Pairwise Comparisons

- Family: ALL PAIRWISE COMPARISON $\mu_i - \mu_k$
- If equal size in all groups ($n_1 = \ldots = n_g = n$), observe that

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

  in which $q = \frac{\bar{y}_{\max} - \bar{y}_{\min}}{\sqrt{\mathrm{MSE}/n}}$ has a ***studentized range distribution***.
- The critical values $q_{g,N-g,\alpha}$ for the studentized range distribution can be found on p.814-815, Table A.8 in the textbook
- Controls the FWER exactly at α if equal group sizes ($n_1 = \ldots = n_g$); no more than α if unequal group sizes

Tukey's Method for Pairwise Comparisons

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

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

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

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

In both the C.I. and the test,

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

Tukey's HSD

If equal group sizes ($n_1 = \ldots = n_g = n$), to be significant at FWER $= \alpha$ based Tukey's correction, the mean difference $\bar{y}_{i\bullet} - \bar{y}_{k\bullet}$ must be at least

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

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

R command to find $q_{g,N-g,\alpha}$: `qtukey(1-alpha,g,N-g)`

```
qtukey(0.95, 6, 18)/sqrt(2)
[1] 3.178
```

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 Method in R

```
library(emmeans)
mod1 = aov(percent ~ trt, data=grass)
mod1emm = emmeans(mod1, "trt")
pairs(mod1emm, infer=c(T,T), level=0.95, adjust="tukey")
```

See the R output on the next page

```
 contrast estimate SE df lower.CL upper.CL t.ratio p.value
 1N - 1Y     12.75  3 18     3.22    22.28   4.253  0.0054
 1N - 2N     13.50  3 18     3.97    23.03   4.503  0.0032
 1N - 3N     26.75  3 18    17.22    36.28   8.924  <.0001
 1N - 4N     44.50  3 18    34.97    54.03  14.845  <.0001
 1N - 4Y     43.00  3 18    33.47    52.53  14.344  <.0001
 1Y - 2N      0.75  3 18    -8.78    10.28   0.250  0.9998
 1Y - 3N     14.00  3 18     4.47    23.53   4.670  0.0022
 1Y - 4N     31.75  3 18    22.22    41.28  10.592  <.0001
 1Y - 4Y     30.25  3 18    20.72    39.78  10.091  <.0001
 2N - 3N     13.25  3 18     3.72    22.78   4.420  0.0038
 2N - 4N     31.00  3 18    21.47    40.53  10.341  <.0001
 2N - 4Y     29.50  3 18    19.97    39.03   9.841  <.0001
 3N - 4N     17.75  3 18     8.22    27.28   5.921  0.0002
 3N - 4Y     16.25  3 18     6.72    25.78   5.421  0.0005
 4N - 4Y     -1.50  3 18   -11.03     8.03  -0.500  0.9955

Confidence level used: 0.95
Conf-level adjustment: tukey method for comparing a family of 6 estimat
P value adjustment: tukey method for comparing a family of 6 estimates
```

Note that the widths of all CIs above are 2x of the HSD.
E.g., the width of the CI for 1Y-1N is $22.28 - 3.22 = 19.06$ is twice of HSD ≈ 9.526.

Summary of Multiple Comparison Adjustments

Method	Family of Tests	Critical Value to Keep FWER $< \alpha$
t-test	a single contrast	$t_{N-g,\alpha/2}$
Tukey	all pairwise comparisons	$q_{g,N-g,\alpha}/\sqrt{2}$
Bonferroni	varies	$t_{N-g,\alpha/(2m)}$, where $m = \#$ of preplanned contrasts
Scheffe	all contrasts	$\sqrt{(g-1)F_{g-1,N-g,\alpha}}$

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