STAT22200 Spring 2014 Chapter 5

Yibi Huang

April 29, 2014

Chapter 5 Multiple Comparisons

Chapter 5 Multiple Comparisons

- Note the t-tests and C.I.'s are constructed assuming we only do one test, or need one confidence interval
- Often we attempt to compare ALL pairs of treatments, as well as contrasts. For an experiment with g treatments, there are
  - $\binom{g}{2} = \frac{r(r-1)}{2}$ pairwise comparisons to make, and
  - numerous contrasts.
- 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.
- It is not fair to hunt around through the data for a big contrast and then pretend that you've only done one comparison. This is called **data snooping**.
  Often the significance of such finding is overstated.

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 *per comparison 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

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

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

Simultaneous Confidence Intervals

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 in the next slide.

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	$\sqrt{(g-1)F_{\alpha,g-1,N-g}}$

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)

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 P-value $< \alpha/k$
- 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 (Bonferoni's Method)

Recall the pairwise comparison for the beet lice example.

Chemical Comparison	Estimate	Standard Error	t-value	p-value	
$\mu_B - \mu_A$	2.960	1.956	1.513	0.13356	> 0.00167
$\mu_C - \mu_A$	6.360	1.956	3.251	0.00159	< 0.00167*
$\mu_D - \mu_A$	12.000	1.956	6.134	1.91×10^{-8}	< 0.00167*
$\mu_C - \mu_B$	3.400	1.956	1.738	0.0854	> 0.00167
$\mu_D - \mu_B$	9.040	1.956	4.621	1.19×10^{-5}	< 0.00167*
$\mu_D - \mu_C$	5.640	1.956	2.883	0.00486	> 0.00167

There are $k = 6$ tests.

For $\alpha = 0.01$, instead of rejecting a null when the P-value $< \alpha$, Bonferoni rejects when the P-value $< \frac{\alpha}{k} = \frac{0.01}{6} = 0.00167$.

Only AC, AD, BD are significantly different.

A B C D

5.4 Tukey's HSD for All Pairwise Comparisons

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

$$
\begin{aligned}
1 - \text{FWER} &= \text{P}(\text{none of the pairwise comparisons is rejected}) \\
&= \text{P}\left(\frac{|\overline{y}_{i\bullet} - \overline{y}_{k\bullet}|}{\sqrt{\text{MSE}\left(\frac{1}{n} + \frac{1}{n}\right)}} < \text{critical value, for all } i \neq k \right) \\
&= \text{P}\left(\frac{\overline{y}_{\max} - \overline{y}_{\min}}{\underbrace{\sqrt{\text{MSE}\left(\frac{1}{n} + \frac{1}{n}\right)}}_{q/\sqrt{2}}} < \text{critical value} \right)
\end{aligned}
$$

in which $q = \dfrac{\overline{y}_{\max} - \overline{y}_{\min}}{\sqrt{\text{MSE}/n}}$, which can be shown to have a **studentized range distribution** for g groups and $N-g$ degrees of freedom, and of which $q_\alpha(g, N-g)$ is the upper $1-\alpha$ quantile.

Setting the critical value at $q_\alpha(g, N-g)/\sqrt{2}$,

$$1 - \text{FWER} = \text{P}(q/\sqrt{2} < q_\alpha(g, N-g)/\sqrt{2}) = 1 - \alpha.$$

the FWER can be controlled at α.

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 the C.I. and the test,

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

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

```
> qtukey(0.95, 4, 25)/sqrt(2)
[1] 2.750643
```

The values of $q_\alpha(g, df)$ are tabulated in Table D.8 on p.633-634 of the textbook.

Example: Beet Lice — Tukey's HSD

```
> beet = read.table("beetlice.txt",header=TRUE)
> aov1 = aov(licecount ~ ttt, data = beet)
> TukeyHSD(aov1, conf.level = 0.99)
  Tukey multiple comparisons of means
    99% family-wise confidence level

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

$ttt
     diff       lwr       upr     p adj
B-A  2.96 -3.294203  9.214203 0.4337634
C-A  6.36  0.105797 12.614203 0.0084911
D-A 12.00  5.745797 18.254203 0.0000001
C-B  3.40 -2.854203  9.654203 0.3099554
D-B  9.04  2.785797 15.294203 0.0000695
D-C  5.64 -0.614203 11.894203 0.0246810
```

Only AC, AD, BD are significantly different.

A B C D

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 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

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

- 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

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

- The critical values for two-sided Dunnett $d_\alpha(g-1, N-g)$ can be found in Table D.9, p.637-638, 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

$$\left| \sum_i a_i b_i \right| \leq \sqrt{\sum_i a_i^2 \sum_i b_i^2}$$

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 use 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

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

Furthermore, 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

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

Example — Beet Lice

Recall

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

, MSE $= 47.84$.

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

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

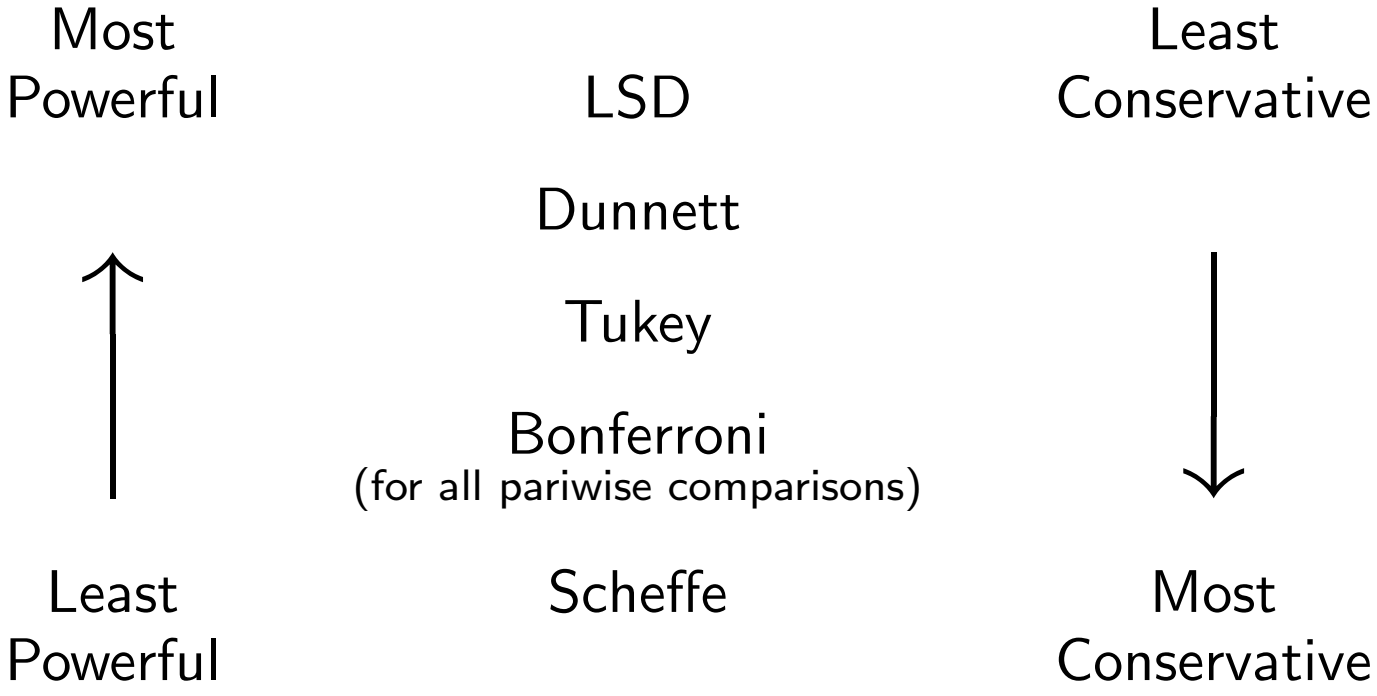
Procedure	LSD	Tukey	Bonferroni	Scheffe
C.I. half width	3.883	5.115	5.270	5.567

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

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.