PS908
Research Methods & Statistics in Psychology

Analysis of Variance (2)
ANOVA Follow-up Tests

Maxwell J Roberts
Department of Psychology
University of Essex
www.tubemapcentral.com

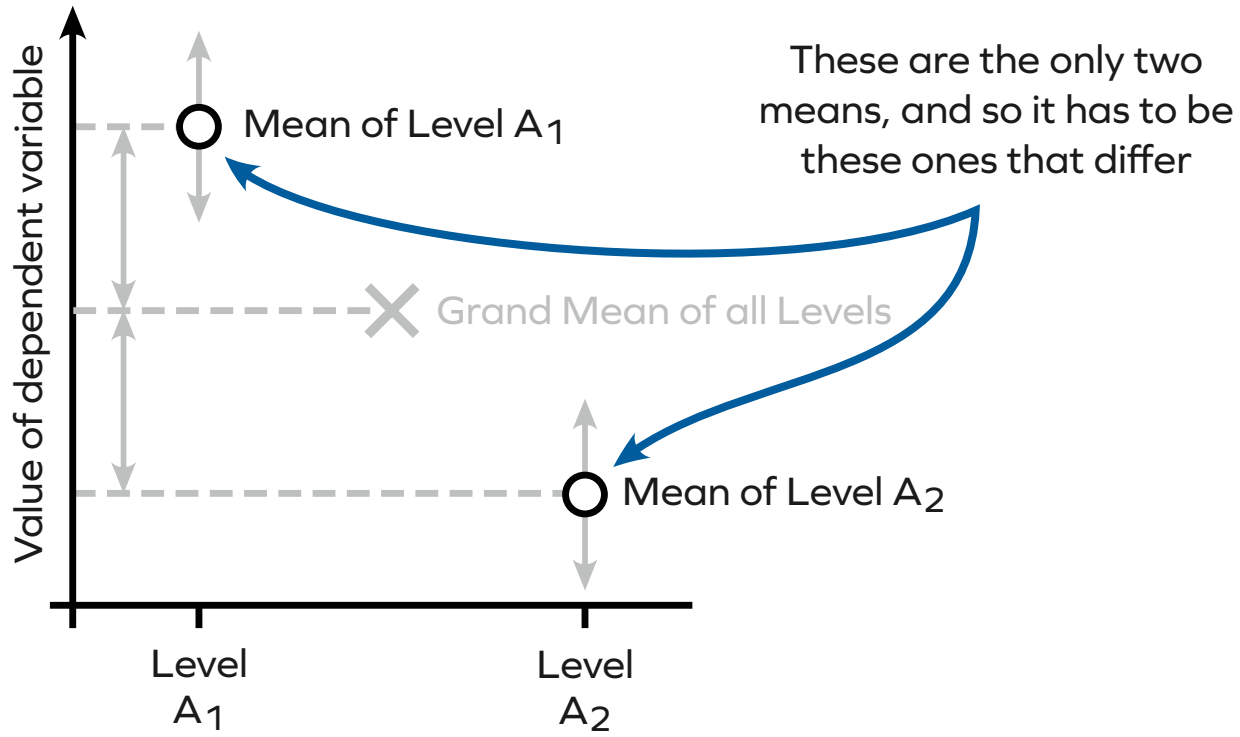
version date: 14/02/2022

Topic 2: ANOVA Follow-Up Tests

- The problem of multiple comparisons
- Strategies for multiple comparisons
- Example experiments
- *Post-hoc* tests (1 & 2): Bonferroni correction, Tukey test
- Planned comparisons
- *Post-hoc* tests (3): Scheffé correction
- Summary: What to do, and when

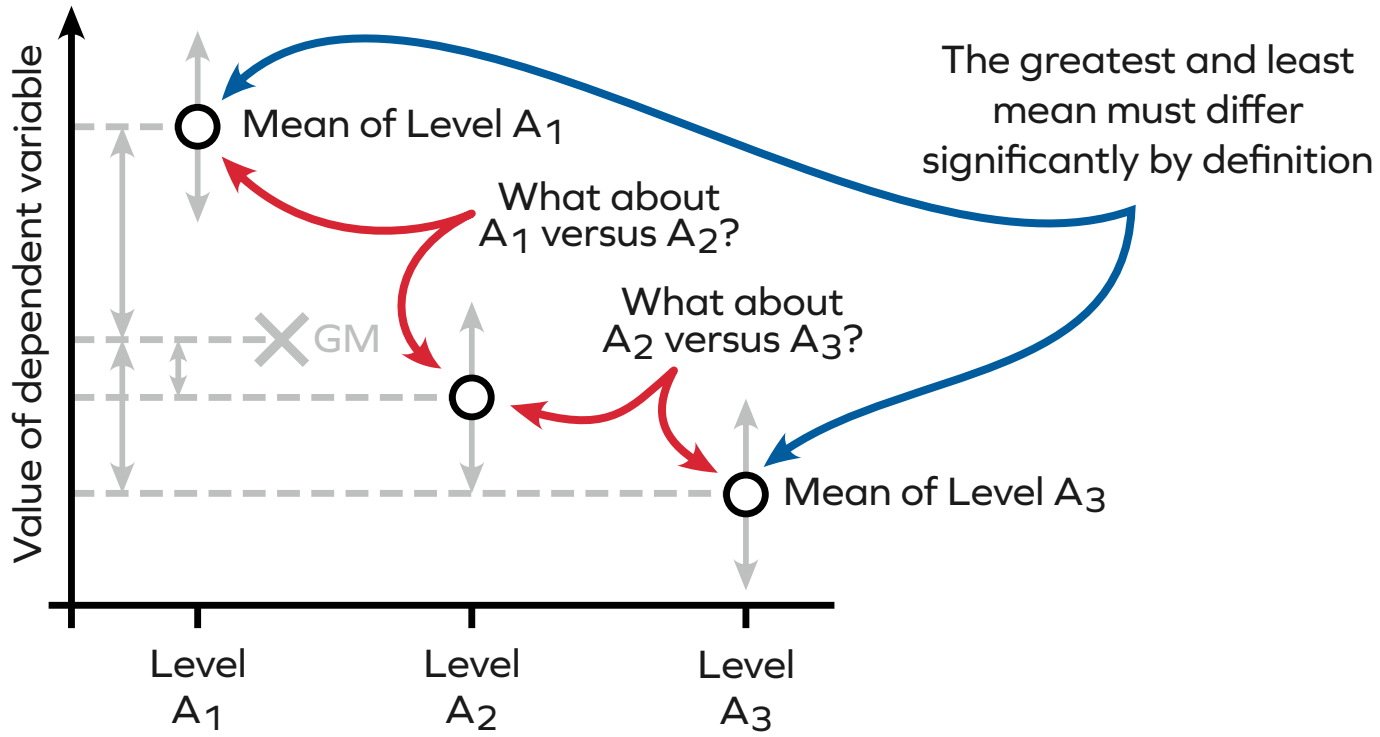
The Problem of Multiple Comparisons

- Two levels in a factor
  - Significant ***difference*** is unambiguous
  - The two means *must* differ significantly



The Problem of Multiple Comparisons

- Three or more levels in a factor
  - Significant ***effect*** is ambiguous
  - Which exact pairs of means differ significantly?



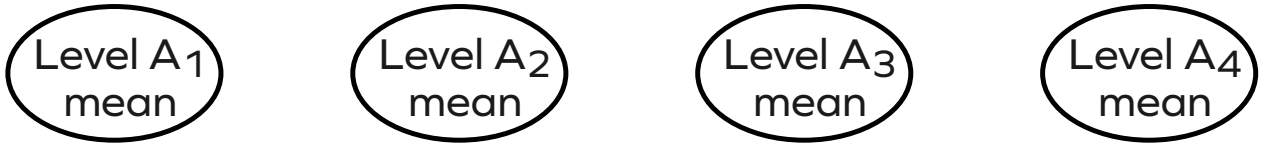
The Problem of Multiple Comparisons

- Why not run extra *pairwise F tests* to follow up?
  - Single factor design, three levels, compare:
    - Level A_1 mean versus Level A_2 mean
    - Lever A_1 mean versus Level A_3 mean
    - Lever A_2 mean versus Level A_3 mean
  - 5% significance level ➔

    Type I Error is made 1 time in 20 on average
  - ***Per-comparison*** Type I Error rate
    - *Probability of a Type I Error for one **single** test*

The Problem of Multiple Comparisons

- Why not run extra *pairwise F tests* to follow up?
  - ***Familywise*** Type I Error rate
    - *Probability of at least Type I Error amongst a **batch** of tests*
  - Single factor design, four levels



The Problem of Multiple Comparisons

- Why not run extra *pairwise F tests* to follow up?
  - Single factor design, four levels
    - Level A_1 mean versus Level A_2 mean
    - Level A_1 mean versus Level A_3 mean
    - Level A_1 mean versus Level A_4 mean
    - Level A_2 mean versus Level A_3 mean
    - Level A_2 mean versus Level A_4 mean
    - Level A_3 mean versus Level A_4 mean
  - Six pairwise comparisons to analyse the data fully
  - Per-comparison Type I Error rate = 5%
  - Familywise Type I Error rate = 26%
    - *The probability of at least one Type I Error in your analysis*

The Problem of Multiple Comparisons

- Why not run extra *pairwise F tests* to follow up?
  - Three factor design, two levels per factor
  - Twelve pairwise comparisons to analyse the data fully
  - Per-comparison Type I Error rate = 5%
  - Familywise Type I Error rate = 46%
    - *The probability of at least one Type I Error in your analysis*
- The problem of the problem of multiple comparisons:
  - *There is no single 'correct' solution!*

Strategies for Multiple Comparisons

1) ***Correct*** the *significance level* of the follow-up test

   Use a more stringent significance level than the chosen significance level, e.g., $p < .05$ ➔ $p < .01$

2) ***Protect*** the *use of* the follow-up test

   Only perform follow-up tests if the overall ANOVA is significant, indicating worthwhile effects to identify

3) ***Select*** from the *potential comparisons* of the follow-up test

   Test only a pre-chosen subset of the possible comparisons

Strategies for Multiple Comparisons

- ***DECIDE IN ADVANCE, BEFORE DATA COLLECTED***

1) No clear predictions ***or***
   every possible comparison necessary/informative

   post-hoc testing

2) Only some of the possible comparisons will be targetted

   planned comparisons

- Four examples of follow-up schemes to be shown

Strategies for Multiple Comparisons

- Four examples of follow-up schemes to be shown
  - *Post-hoc testing* (1):
    using ***pairwise F tests*** and the ***Bonferroni correction***
  - *Post-hoc testing* (2):
    using the ***Tukey test***
  - *Planned comparisons*:
    using ***pairwise F* tests** for small versus large numbers of comparisons
  - *Post-hoc testing* (3):
    using ***pairwise F* tests** and the ***Scheffé correction***

Example Experiment (a)

	Lectures + app	Lectures + tutorials	Lectures only
Level Means	63.4	64.0	59.0

Dependent Variable: end of year module %

- 3 levels in the factor: **a = 3**
- 10 scores per level: **s = 10**

Example Experiment (a)

Source	Sum of Squares	Degrees of Freedom	Variance (Mean Square)	*F*-value	*p*-value (sig. level)
A BETWEEN-GROUP	149.1	2	74.5	3.31	$p > .05$ NS
S/A WITHIN-GROUP	608.4	27	22.5		
TOTAL	757.5	29			

- Obtained value: $F(2,27) = 3.31$
- Critical value: $F_{CRIT} = 3.35$
- Insufficient evidence for treatment effects

Example Experiment (b)

	Placebo	Drug 1	Drug 2	Drug 3
Level Means	550	562	653	601

Dependent Variable: mean reaction time in milliseconds

- 4 levels in the factor: **a = 4**
- 10 scores per level: **s = 10**

Example Experiment (b)

Source	Sum of Squares	Degrees of Freedom	Variance (Mean Square)	*F*-value	*p*-value (sig. level)
A BETWEEN-GROUP	**9600**	**3**	**3200**	**10.2**	$p < .01$ **
S/A WITHIN-GROUP	**11282.4**	**36**	**313.4**		
TOTAL	**20882.4**	**39**			

- Obtained value: $F(3,36) = 10.2$
- Critical value: $F_{CRIT} = 2.87$
- Good statistical evidence for treatment effects

Post-Hoc Tests (1)

- No clear predictions ***or***
  every possible comparison necessary/informative

  1) *Post-hoc* comparisons using ***pairwise F tests***
     and the ***Bonferroni correction***

Correct: stringent, full correction to critical value

~~**Protect:**~~ fully corrected, no need for protection

~~**Select:**~~ all possible comparisons intended to be made

Post-Hoc Tests (1)

- No need for significant *F* value from main ANOVA table
- Use additional *F* tests to make all possible pairwise comparisons between means
- New significance level = original significance level divided by the number of comparisons
- *Fully corrected*, never worry about familywise Type I Error ever again
- Rather a stringent correction
- Lacks *statistical power*

Post-Hoc Tests (1)

- Applying Bonferroni correction to the statistics app experiment:

	Lectures + app	Lectures + tutorials	Lectures only
Level Means	63.4	64.0	59.0

Dependent Variable: end of year module %

Source	Sum of Squares	Degrees of Freedom	Variance (Mean Square)	*F*-value	*p*-value (sig. level)
A BETWEEN-GROUP	149.1	2	74.5	3.31	$p > .05$ NS
S/A WITHIN-GROUP	608.4	27	22.5		
TOTAL	757.5	29			

Post-Hoc Tests (1)

- Quick formula for pairwise comparisons once main ANOVA table is computed

$$F = \frac{s/2\,(\overline{A}_i - \overline{A}_j)^2}{\text{Variance}_{\text{ERROR}}}$$

- From statistics app experiment
  - **s** = number of scores in each level = **10**
  - Error term is from main ANOVA table = **22.5**
  - **df = 1** for between-group variance; **(a – 1)**
    df = 27 for the error term (same as original table)

Post-Hoc Tests (1)

	Lectures + app	Lectures + tutorials	Lectures only
Level Means	63.4	64.0	59.0

Dependent Variable: end of year module %

$$F = \frac{\frac{s}{2}(\overline{A}_i - \overline{A}_j)^2}{\text{Variance}_{\text{ERROR}}} = \frac{5(\overline{A}_i - \overline{A}_j)^2}{22.5}$$

- *L+app* vs. *L-only* group

Post-Hoc Tests (1)

	Lectures + app	Lectures + tutorials	Lectures only
Level Means	63.4	64.0	59.0

Dependent Variable: end of year module %

$$F = \frac{\frac{s}{2}(\overline{A}_i - \overline{A}_j)^2}{\text{Variance}_{\text{ERROR}}} = \frac{5(\overline{A}_i - \overline{A}_j)^2}{22.5}$$

- *L+app* vs. *L+tutorial* group

Post-Hoc Tests (1)

	Lectures + app	Lectures + tutorials	Lectures only
Level Means	63.4	64.0	59.0

Dependent Variable: end of year module %

$$F = \frac{\frac{s}{2}(\overline{A}_i - \overline{A}_j)^2}{\text{Variance}_{\text{ERROR}}} = \frac{5(\overline{A}_i - \overline{A}_j)^2}{22.5}$$

- *L+tutorial* vs. *L-only* group

Post-Hoc Tests (1)

- Obtained values:
  - *L+app* vs. *L-only: F*(1,27) = 4.30
  - *L+app* vs. *L+tutorial: F*(1,27) = 0.08
  - *L+tutorial* vs. *L-only: F*(1,27) = 5.56
- Critical value for df = (1,27) at **.05** sig. level: $F_{\mathbf{CRIT}}$ = **4.21**

- ***BUT*** we have made three comparisons
- ***Bonferroni correction:***

  [.05 significance level] ÷ [3 pairwise comparisons] = .017

Post-Hoc Tests (1)

- Obtained values:
  - *L+app* vs. *L-only:* $F(1,27) = 4.30$
  - *L+app* vs. *L+tutorial:* $F(1,27) = 0.08$
  - *L+tutorial* vs. *L-only:* $F(1,27) = 5.56$
- Critical value for df = (1,27) at **.017** sig. level: $F_{CRIT} = \mathbf{6.47}$

- No significant differences between teaching methods

- Bonferroni correction has 'removed' two potential effects

Post-Hoc Tests (2)

- No clear predictions ***or***
  every possible comparison necessary/informative

  2) *Post-hoc* comparisons
     using the ***Tukey test***

✓ **Correct:** but less stringent than Bonferroni correction

✓ **Protect:** ANOVA must be significant before proceeding

✗ ~~**Select:**~~ all possible comparisons intended to be made

[Newman-Keuls is similar, less stringent than Tukey]

Post-Hoc Tests (2)

- Need significant *F* value from main ANOVA table
- Test is designed to make all possible pairwise comparisons between means
- Special formula corrects according to number of comparisons and gives a ***critical difference***
- Less stringent than Bonferroni correction once the initial ANOVA protection has been satisfied

Post-Hoc Tests (2)

	Lectures + app	Lectures + tutorials	Lectures only
Level Means	63.4	64.0	59.0

Dependent Variable: end of year module %

Source	Sum of Squares	Degrees of Freedom	Variance (Mean Square)	*F*-value	*p*-value (sig. level)
A BETWEEN-GROUP	**149.1**	**2**	**74.5**	**3.31**	$p > .05$ NS
S/A WITHIN-GROUP	**608.4**	**27**	**22.5**		
TOTAL	**757.5**	**29**			

- *F* on main ANOVA table is non-significant

Post-Hoc Tests (2)

- No evidence for treatment effects
- Tukey test is protected: cannot proceed further
- Differences in means likely to be due to experimental error
- No other tests possible

Post-Hoc Tests (2)

	Placebo	Drug 1	Drug 2	Drug 3
Level Means	550	562	653	601

Dependent Variable: mean reaction time in milliseconds

Source	Sum of Squares	Degrees of Freedom	Variance (Mean Square)	*F*-value	*p*-value (sig. level)
A BETWEEN-GROUP	**9600**	**3**	**3200**	**10.2**	$p < .01$ **
S/A WITHIN-GROUP	**11282.4**	**36**	**313.4**		
TOTAL	**20882.4**	**39**			

- *F* on main ANOVA table is significant
- Sound evidence for treatment effects
- Apply Tukey test to investigate the effects

Post-Hoc Tests (2)

$$\text{Critical difference, } W = r\sqrt{\frac{\text{Variance}_{\text{ERROR}}}{s}}$$

- *r* is obtained from *studentised range statistic* tables
  - Depends on number of levels ($a = 4$) and df_{ERROR} (36)
  - $r = 3.84$ for this experiment
  - You will never need to use the table/equation, ever!

$$W = 3.84\sqrt{\frac{313.4}{10}} = 21.5$$

Post-Hoc Tests (2)

	Placebo	Drug 1	Drug 2	Drug 3
Level Means	550	562	653	601

- Critical difference = 21.5
  - Placebo vs. Drug 1
  - Placebo vs. Drug 2
  - Placebo vs. Drug 3
  - Drug 1 vs. Drug 2
  - Drug 1 vs. Drug 3
  - Drug 2 vs. Drug 3

Post-Hoc Tests (2)

	Placebo 550	Drug 1 562	Drug 3 601	Drug 2 653
Placebo	—	NS	*	*
Drug 1		—	*	*
Drug 3			—	*
Drug 2				—

- Computers often show output as a matrix

- Results show that ***ONLY*** Drug 1 is safe to take
- All other drugs significantly reduce performance compared with both the Placebo and also Drug 1

Planned Comparisons

- Only some of the possible comparisons will be targetted

3) Planned comparisons using ***pairwise F* tests** and (*possibly*) the ***Bonferroni correction***

- Increase statistical power by thinking ahead
- Use pairwise *F* tests to make only the ***crucial*** comparisons for a hypotheses
- Advance planning/honesty/economy means no need to protect or correct the tests

Planned Comparisons

- **Few comparisons**
  - **(a – 1)** comparisons [number of levels minus 1]

✗ ~~**Correct:**~~ no need, Type I Error is kept under control

✗ ~~**Protect:**~~ no need, Type I Error is kept under control

✓ **Select: (a – 1)** comparisons

  - Very powerful, always use this method if possible

Planned Comparisons

- **Many comparisons**

  - More than **(a – 1)** comparisons (but not everything)

  ✓ **Correct:** Apply ***Bonferroni*** for number of comparisons

  ✗ ~~**Protect:**~~ no need, Type I Error is kept under control

  ✓ **Select:** More than **(a – 1)** comparisons

  - Less powerful, but still better than testing all possible comparisons, i.e. *post-hoc* tests

Planned Comparisons

	Lectures + app	Lectures + tutorials	Lectures only
Level Means	63.4	64.0	59.0

- ***What is the research question?***
  - What is the best teaching method, what is the worst teaching method?
    - Must test ***all three pairwise comparisons***
  - Is it worth adopting the app?
    - Only need to test ***two of the pairwise comparisons***

Planned Comparisons

	Lectures + app	Lectures + tutorials	Lectures only
Level Means	63.4	64.0	59.0

- ***What is the research question?***
  - Is it worth adopting the app?
    - Only need to test ***two of the pairwise comparisons***
    - ***Before*** collecting data, planned to compare
      - *L+app* vs. *L-only* group
      - *L+app* vs. *L+tutorial* group

Planned Comparisons

- Same quick formula for pairwise comparisons once main ANOVA table is computed

$$F = \frac{\frac{s}{2}(\overline{A}_i - \overline{A}_j)^2}{\text{Variance}_{\text{ERROR}}}$$

- Obtained values:
  - *L+app* vs. *L-only:* $F(1,27) = 4.30$
  - *L+app* vs. *L+tutorial:* $F(1,27) = 0.08$
- Critical value for df = (1,27) at **.05** sig. level: $F_{\text{CRIT}} = \mathbf{4.21}$
- **(a – 1)** comparisons, so no need for Bonferroni correction

Planned Comparisons

	Lectures + app	Lectures + tutorials	Lectures only
Level Means	63.4	64.0	59.0

Dependent Variable: end of year module %

- Obtained values:
  - *L+app* vs. *L-only:* $F(1,27) = 4.30$ $p < .05$, Sig
  - *L+app* vs. *L+tutorial:* $F(1,27) = 0.08$ $p > .05$, NS

- The app ***is*** significantly better than no assistance, but is ***not*** significantly better than tutorial support

Post-Hoc Tests (3)

- Have performed planned comparisons ***but*** the unplanned comparisons look interesting

  4) *Post-hoc* comparisons using ***pairwise F* tests** and the ***Scheffé correction***

✓ **Correct:** most stringent correction to the critical value

✗ ~~**Protect:**~~ more than fully corrected, no protection

✗ ~~**Select:**~~ any comparisons can be tested

Post-Hoc Tests (3)

- No need for significant *F* value from main ANOVA table
- Use additional *F* tests to make any further pairwise comparisons between means
- New corrected critical value obtained from formula
- Very stringent correction
- Use if exploration after planned comparisons is desired

Post-Hoc Tests (3)

	Lectures + app	Lectures + tutorials	Lectures only
Level Means	63.4	64.0	59.0

- ***Before*** collecting data, planned to compare
  - *L+app* vs. *L-only* group
  - *L+app* vs. *L+tutorial* group
- Two of the three possible comparisons have already been planned and performed
  - But the *L+tutorial* vs. *L-only* group comparison looks interesting, the largest difference of all

Post-Hoc Tests (3)

- Same quick formula for pairwise comparisons once main ANOVA table is computed

$$F = \frac{\frac{s}{2}(\overline{A}_i - \overline{A}_j)^2}{\text{Variance}_{\text{ERROR}}}$$

- Obtained values:
  - *L+tutorial* vs. *L-only:* $F(1,27) = 5.56$
- Critical value for df = (1,27) at **.05** sig. level: $F_{\text{CRIT}} = \mathbf{4.21}$
- Analysis allowance used up, ***Scheffé correction*** needed

Post-Hoc Tests (3)

Source	Sum of Squares	Degrees of Freedom	Variance (Mean Square)	*F*-value	*p*-value (sig. level)
A BETWEEN-GROUP	149.1	2	74.5	3.31	$p > .05$ NS
S/A WITHIN-GROUP	608.4	27	22.5		
TOTAL	757.5	29			

- Obtained value: $F(2,27) = 3.31$
- Critical value: $F_{\mathbf{CRIT}} = 3.35$

$$F_{\text{CRIT\_SCHEFFÉ}} = (a-1)F_{\text{CRIT\_MAIN\_TABLE}}$$

Post-Hoc Tests (3)

- Obtained value:
  - *L+tutorial* vs. *L-only:* $F(1,27) = 5.56$
- Critical value: $F_{CRIT\_SCHEFFÉ}$ = **6.70**

- No evidence that tutorials are significantly better than no support, despite the highest difference between means

- There is payback for lenient planned comparisons:
  - Uncorrected F_{CRIT} for planned comparisons = 4.21
  - Corrected $F_{CRIT\_BONFERRONI}$ for ***all*** *post-hoc* tests = 6.47
  - Corrected $F_{CRIT\_SCHEFFÉ}$ for ***all+*** *post-hoc* tests = 6.70

Summary: What to Do, and When

- *Post-hoc* tests (if all comparisons must be made), in order of *stringency* of significance level *correction*:
  - Bonferroni correction: most stringent/least power
  - Tukey test: less stringent/more power
  - [Newman-Keuls test: least stringent/most power]

  - Choice depends on the relative damage of a Type I Error versus a Type II Error

Summary: What to Do, and When

- Planned comparisons (do these if possible)
- **ALWAYS** in advance, ***NEVER*** if you have seen the data
  - **(a – 1)** comparisons: no correction necessary
  - **> (a – 1)** comparisons: apply Bonferroni correction

  - The more economy the better
  - Adding levels to a factor ***loses*** statistical power
  - Make your larger design work
    for you: analyse it conscientiously

Summary: What to Do, and When

- Additional tests (after planned comparisons used up)
  - Apply Scheffé correction
  - Beware, different levels of statistical power in the same data analysis can be disorientating

Summary: What to Do, and When

- Decide on your analysis scheme ***BEFORE*** collecting data
- ***DO NOT*** try lots of different schemes on the same data trying to find the one that you like the best
- Issues are less important when analysing large datasets: collect more data to prevent subsequent statistics agony