Simultaneous Comparisons and the Control of Type I Errors
Chapter 6

ERSH 8310

Lecture 5

September 16, 2009

Today's Class

- Simultaneous comparisons
  - Controlling Type-I Error Rates

- The word of the day: Meh

RESEARCH QUESTIONS AND TYPE I ERROR

Research Questions and Type I Error

- This chapter examines the problem of *cumulative Type I errors* and the solutions designed to avoid them
- Researchers are often interested in a set of related hypothesis (i.e., a family of tests)
- The per-comparison error, called α, uses each comparison as the conceptual unit for determining Type I error
- The family-wise (FW) Type I error, denoted as α_{FW}, considers the probability of making one or more Type I errors in the set of comparisons under scrutiny.

Relationship Between Both Kinds of Type I Error

- The relationship between the two kinds of Type I error is:

$$\alpha_{FW} = 1 - (1 - \alpha)^c$$

- *c* represents the number of orthogonal comparisons that are conducted

- How this works is to consider each hypothesis test independently
  - Pr(No Familywise Errors) = $(1 - \alpha)\,(1 - \alpha)\ldots(1 - \alpha) = (1 - \alpha)^c$

- An approximate calculation is: $\tilde{\alpha}_{FW} = c\alpha$
  - We will use this later

What Did That Mean???

- Consider an experiment with four treatment levels
  - Our vigilance task example, for instance

Then:

- If you set the overall Type-I error rate to be 0.05
- And you tested the difference between each pairing of means (6 pairs total)
- Then the $\alpha_{FW} = 1-(1-.05)^6 = 0.264$
- This means you would have a 26.4% chance of making a Type I error somewhere in your experiment

General Plans for Experiments

- There are three general plans of an experiments:

1. Testing the primary questions
   - Do the treatment means differ generally?

2. Looking at special families of hypotheses
   - Contrasts
   - Tests for linear trends
   - Planned comparisons

3. Exploring the data for unexpected relationships
   - Unplanned post-hoc tests

PLANNED COMPARISONS

Planned Comparisons

- Experiments can be designed with specific hypotheses in mind without reference to the outcome of the omnibus F test
  - ➢ The most widely used strategy to control the family-wise error rate is to evaluate the planned comparisons in a normal way (e.g., α)

- The value of orthogonal comparisons lies in the independence of inference

Planned Comparisons

- Meaningful comparisons may contain some nonorthogonal comparisons
  - The nonorthogonal comparisons should be interpreted with particular care

- One may limit the number of planned comparisons (e.g., the number may be df_A = A-1)
  - Many researchers limit the number of planned comparisons depending on the research hypotheses and on the complexity of the experiment

RESTRICTED SETS OF CONTRASTS

Restricted Sets of Contrasts

- If you have a plan for the number of contrasts you would like to make a priori, then the following procedures can help adjust your overall Type-I error rate so that you have more protection from error:
    - Bonferroni
    - Sidák-Bonferroni
    - Dunnett's Test

- Any of these tests will help in making decisions when the number of hypothesis tests is known prior to the experiment

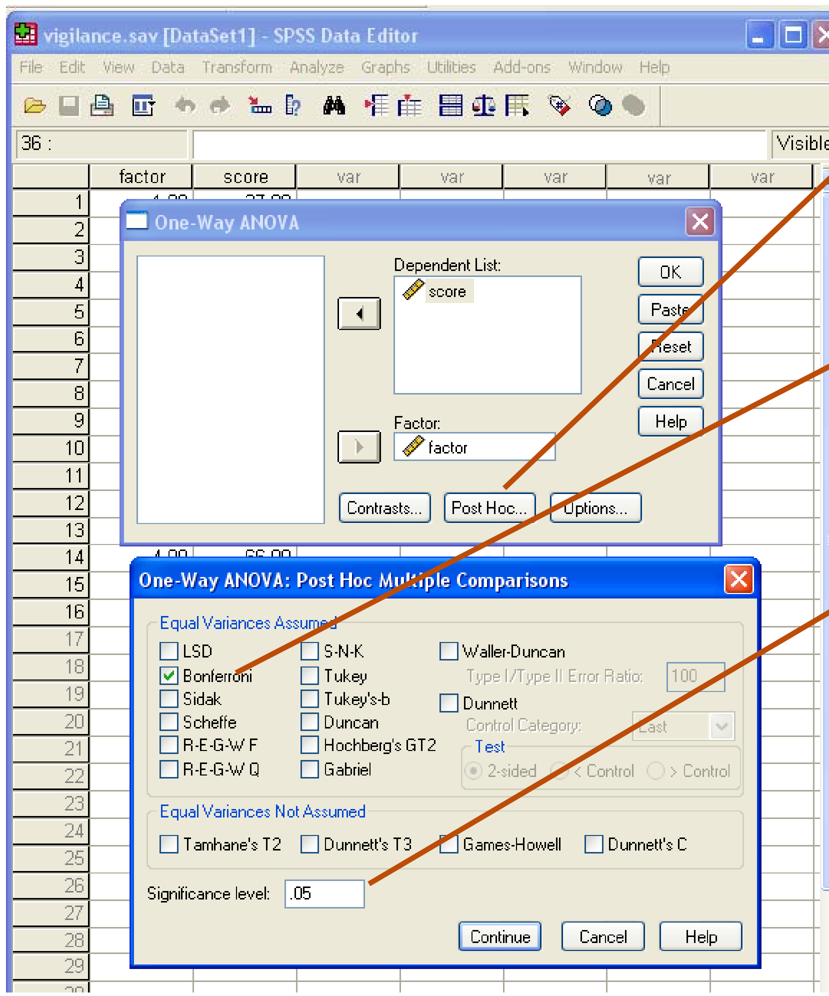
The Bonferroni Procedure

- We may apply some corrections to control the overall error rate
- The Bonferroni correction is the most widely applicable family wise control procedure for small families
- Because $\tilde{\alpha}_{FW} = c\alpha$ we may use the Bonferroni test or the Dunnet Test that uses:

$$\alpha = \frac{\alpha_{FW}}{c}$$

Where a is the new percomparison significance level and c is the number of comparisons

Bonferroni Example – SPSS Steps



Under the Post Hoc…Box

Check Bonferroni

Set your significance level (Type I error or α)

Bonferroni Example – SPSS Output

Post Hoc Tests

This tells us the means are significantly different for levels 1 and 3, and 1 and 4.

Multiple Comparisons

Dependent Variable: score
Bonferroni

(I) factor	(J) factor	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound
1.00	2.00	-11.25000	8.67347	1.000	-38.5947	16.0947
	3.00	-31.00000*	8.67347	.023	-58.3447	-3.6553
	4.00	-35.25000*	8.67347	.009	-62.5947	-7.9053
2.00	1.00	11.25000	8.67347	1.000	-16.0947	38.5947
	3.00	-19.75000	8.67347	.251	-47.0947	7.5947
	4.00	-24.00000	8.67347	.102	-51.3447	3.3447
3.00	1.00	31.00000*	8.67347	.023	3.6553	58.3447
	2.00	19.75000	8.67347	.251	-7.5947	47.0947
	4.00	-4.25000	8.67347	1.000	-31.5947	23.0947
4.00	1.00	35.25000*	8.67347	.009	7.9053	62.5947
	2.00	24.00000	8.67347	.102	-3.3447	51.3447
	3.00	4.25000	8.67347	1.000	-23.0947	31.5947

*. The mean difference is significant at the .05 level.

The Sidák-Bonferroni Procedure

- This procedure uses:

$$\alpha = 1 - \left(1 - \alpha_{FW}\right)^{\frac{1}{c}}$$

Which is the exact level (as opposed to the approximate given in the Bonferroni test).

Dunnett's Test

- The Dunnett's test is a specialized family-wise correction technique that compensates for the increased number of potential Type I errors that involves only the control-experimental contrast

- It is relevant to all pairwise comparisons involving a single group

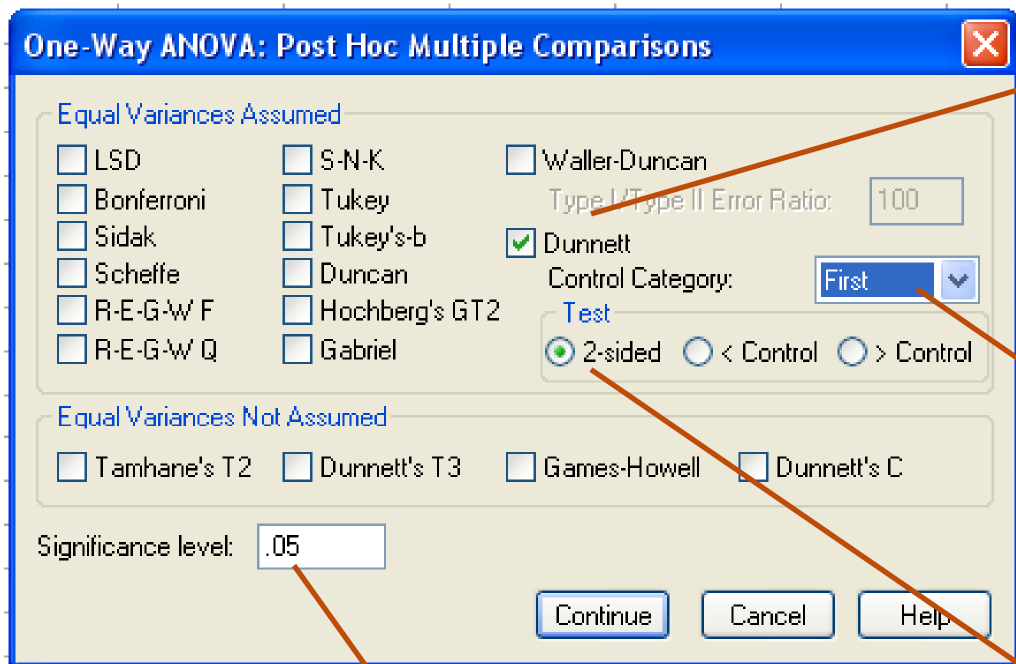
- The critical values of t (i.e., $t_{Dunnett}$) are presented in Appendix A.5 (pp. 582-585)

Dunnett’s Test: When To Use

- Dunnett’s test is more powerful (will be able to detect mean differences better) than either the Bonferroni or the S-B procedures

- It typically is used whenever one group (most commonly the control group) is being compared to all the other a-1 groups (most commonly the experimental groups)

Dunnett Example: SPSS Steps



Under Post Hoc, select the Dunnett check box.

Pick the category for the control group.

Pick the type of test: 2-sided is just for any difference, the others are directional hypotheses.

Set your significance level (Type I error or α)

Dunnett Example: SPSS Output

Post Hoc Tests

Multiple Comparisons

Dependent Variable: score
Dunnett t (2-sided)[a]

(I) factor	(J) factor	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
2.00	1.00	11.25000	8.67347	.453	-12.0199	34.5199
3.00	1.00	31.00000*	8.67347	.010	7.7301	54.2699
4.00	1.00	35.25000*	8.67347	.004	11.9801	58.5199

*. The mean difference is significant at the .05 level.

a. Dunnett t-tests treat one group as a control, and compare all other groups against it.

PAIRWISE COMPARISONS

Pairwise Comparisons

- Pairwise comparisons are used for looking at all possible pairings of treatment means
  - They protect you from making more Type I errors by making the threshold for significant mean differences larger
- We will discuss three methods: Tukey, Fisher-Hayter, and Newman-Keuls
  - For other methods, see Seaman, Levine, and Serlin (1991) or Toothaker (1991)
- The Tukey (1953) procedure may be used to maintain the family-wise rate at the chosen value of α_{FW} for the entire set of pairwise comparisons

Tukey's HSD Procedure

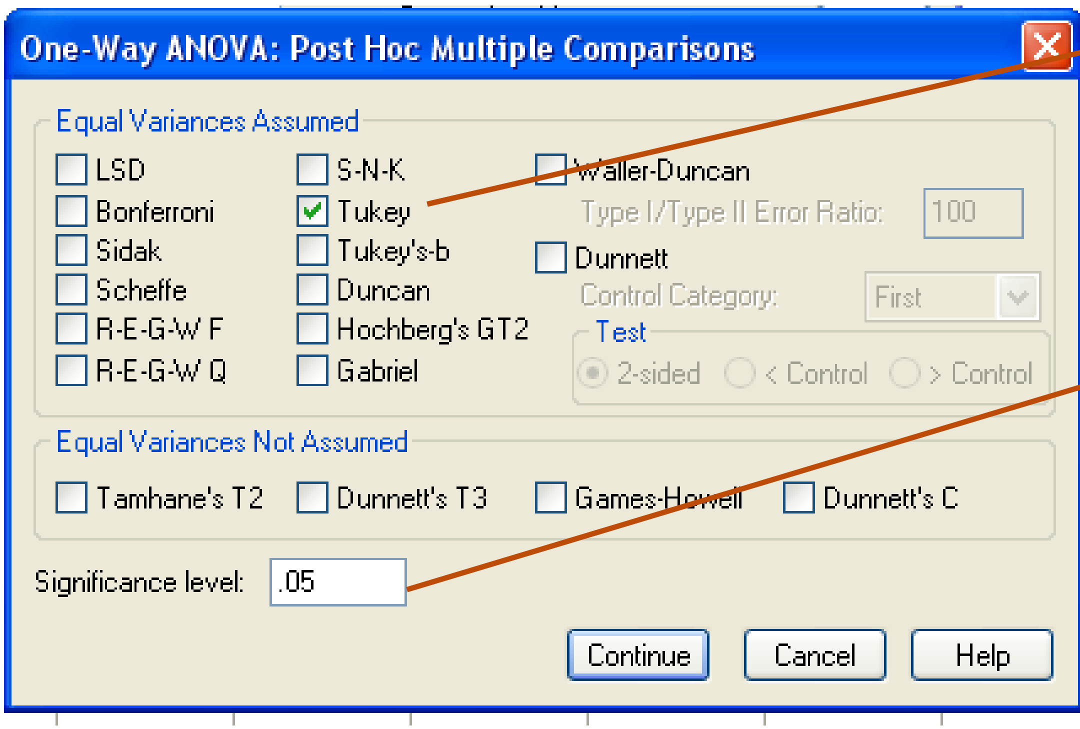
- The pairwise difference between means must exceed the critical value:

$$D_{Tukey} = q_a s_M = q_a \sqrt{\frac{MS_{S/A}}{n}}$$

where q_a is an entry in Appendix A.6 (see pp. 586-589).

Note the there exists a different critical difference for the variance heterogeneity case (see Equation 6.8)

Tukey Example: SPSS Steps



Under Post Hoc, select the Tukey check box.

Set your significance level (Type I error or α)

Tukey Example: SPSS Output (Part 1)

Post Hoc Tests

Multiple Comparisons

Dependent Variable: score
Tukey HSD

(I) factor	(J) factor	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
1.00	2.00	-11.25000	8.67347	.582	-37.0007	14.5007
	3.00	-31.00000*	8.67347	.017	-56.7507	-5.2493
	4.00	-35.25000*	8.67347	.007	-61.0007	-9.4993
2.00	1.00	11.25000	8.67347	.582	-14.5007	37.0007
	3.00	-19.75000	8.67347	.158	-45.5007	6.0007
	4.00	-24.00000	8.67347	.071	-49.7507	1.7507
3.00	1.00	31.00000*	8.67347	.017	5.2493	56.7507
	2.00	19.75000	8.67347	.158	-6.0007	45.5007
	4.00	-4.25000	8.67347	.960	-30.0007	21.5007
4.00	1.00	35.25000*	8.67347	.007	9.4993	61.0007
	2.00	24.00000	8.67347	.071	-1.7507	49.7507
	3.00	4.25000	8.67347	.960	-21.5007	30.0007

*. The mean difference is significant at the .05 level.

Tukey Example: SPSS Output (Part 2)

Homogeneous Subsets

score

Tukey HSD[a]

factor	N	Subset for alpha = .05	
		1	2
1.00	4	26.5000	
2.00	4	37.7500	37.7500
3.00	4		57.5000
4.00	4		61.7500
Sig.		.582	.071

Means for groups in homogeneous subsets are displayed.

a. Uses Harmonic Mean Sample Size = 4.000.

This displays the groups of means that are not significantly different from each other.

Here, 1 and 2 are not different and 2, 3, and 4 are not different.

The Fisher-Hayter Procedure

- Several other procedures have been developed to increase the power of the test
- The Fisher-Hayter procedure uses a sequential approach to testing and involves two steps
- Conduct an omnibus test at a_{FW} level
  - If it is significant, then go to the treatment means
- Test all pairwise comparisons using the critical difference:

$$D_{FH} = q_{a-1}\sqrt{\frac{MS_{S/A}}{n}}$$

- Note: not in SPSS

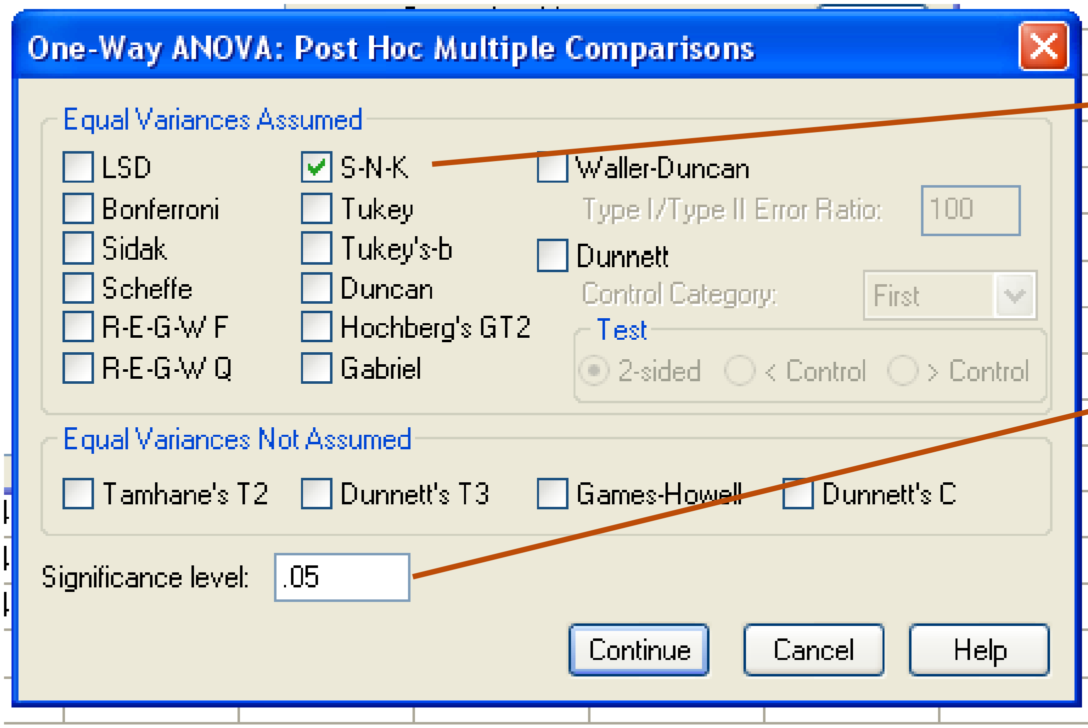
The Newman-Keuls and Related Procedures

- The critical difference is given by:

$$D_{NKk} = q_k \sqrt{\frac{MS_{S/A}}{n}}$$

- where k = a initially and declines until the largest difference becomes not significant

NK Example: SPSS Steps



Under Post Hoc, select the S-N-L check box.

Set your significance level (Type I error or α)

NK Example: SPSS Output

Homogeneous Subsets

score

Student-Newman-Keuls[a]

factor	N	Subset for alpha = .05	
		1	2
1.00	4	26.5000	
2.00	4	37.7500	
3.00	4		57.5000
4.00	4		61.7500
Sig.		.219	.633

Means for groups in homogeneous subsets are displayed.

a. Uses Harmonic Mean Sample Size = 4.000.

Notice anything different from the Tukey procedure?

Recommendations from the Book

- The process of pairwise comparisons is typically the same, regardless of which test you usr
  - Look at a bunch of p-values…determine which means are different
  - Meh

- The tests differ in the degree of conservativeness each may present

- The book recommends using either Tukey's procedure or the Fisher-Hayter procedure
  - Meh

POST HOC ERROR CORRECTION

Post Hoc Error Correction

- Fisher's (1935) procedure (i.e., to test the omnibus F, followed by the unrestricted testing of comparisons among the means, if and only if the overall F is significant), called the least significant difference test, controls the family-wise error indirectly

- This procedure has been criticized by many for not providing adequate control over the family-wise error

- There are several alpha-adjusted techniques
  - We will consider the procedure by Scheffé

Scheffé's Procedure

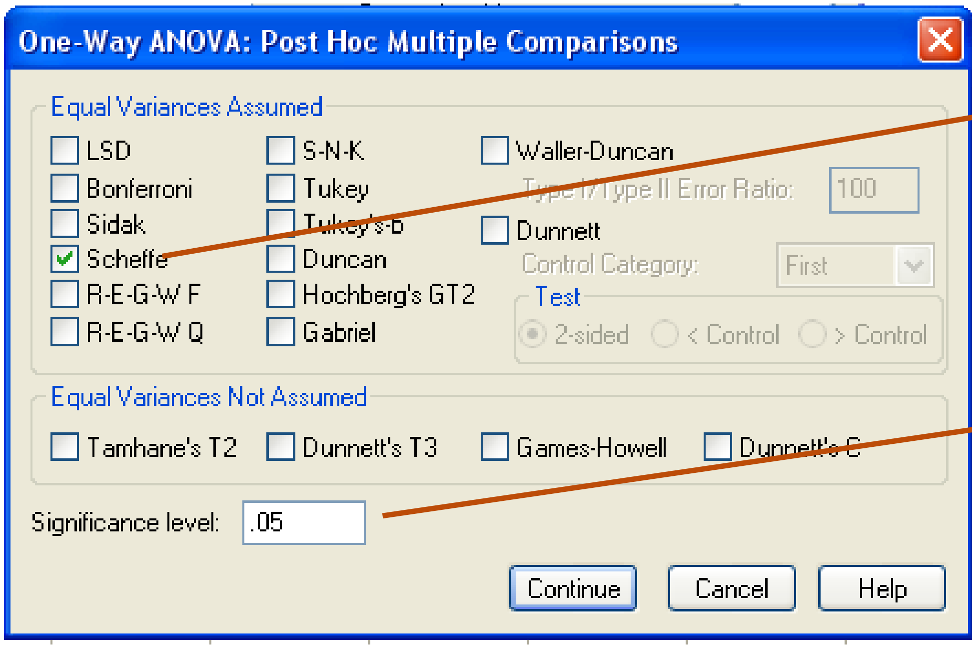
- Scheffé's (1953) procedure is a technique that allows a researcher to maintain the family-wise rate at a particular value regardless of the number of comparisons actually conducted

- The critical value is

$$F_{Scheffe} = (a-1)F_{\alpha_{EW}}(df_A, df_{S/A})$$

Where α_{EW} is the experiment wise error rate (see p. 112)

Scheffé Example: SPSS Steps



Under Post Hoc, select the Scheffe check box

Set your significance level (Type I error or α)

Scheffé Example: SPSS Output (Part 1)

Post Hoc Tests

Multiple Comparisons

Dependent Variable: score
Scheffe

(I) factor	(J) factor	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
1.00	2.00	-11.25000	8.67347	.651	-39.3163	16.8163
	3.00	-31.00000*	8.67347	.029	-59.0663	-2.9337
	4.00	-35.25000*	8.67347	.013	-63.3163	-7.1837
2.00	1.00	11.25000	8.67347	.651	-16.8163	39.3163
	3.00	-19.75000	8.67347	.214	-47.8163	8.3163
	4.00	-24.00000	8.67347	.105	-52.0663	4.0663
3.00	1.00	31.00000*	8.67347	.029	2.9337	59.0663
	2.00	19.75000	8.67347	.214	-8.3163	47.8163
	4.00	-4.25000	8.67347	.970	-32.3163	23.8163
4.00	1.00	35.25000*	8.67347	.013	7.1837	63.3163
	2.00	24.00000	8.67347	.105	-4.0663	52.0663
	3.00	4.25000	8.67347	.970	-23.8163	32.3163

*. The mean difference is significant at the .05 level.

Scheffé Example: SPSS Output (Part 2)

Homogeneous Subsets

score

Scheffe[a]

factor	N	Subset for alpha = .05	
		1	2
1.00	4	26.5000	
2.00	4	37.7500	37.7500
3.00	4		57.5000
4.00	4		61.7500
Sig.		.651	.105

Means for groups in homogeneous subsets are displayed.

a. Uses Harmonic Mean Sample Size = 4.000.

Wrapping Up…

- The ANOVA procedure yields an omnibus F test that tells you that at least one group mean is different from the rest
- Today we talked about ways in which you could find out which pair(s) of means happened to be different
- Simultaneous comparisons are specific hypothesis tests that examine how each mean may differ from all the other means
- By using any of the methods described today, we protect ourselves from making Type-I errors in our studies

Up Next…

- In Lab:
  - How to do post-hoc tests in SPSS
  - Example data
- Homework:
  - Posted now – due at beginning of class next week
- Next week reading:
  - Read Chapter 7 – linear models