Chapter 6
Planned Contrasts and Post-hoc Tests for one-way ANOVA

Planned Contrasts and Post-hoc Tests
for one-way ANOVA

First, a cautionary note about playing the p-value game.

1. The problem of multiple contrasts

   - Contrasts give you the flexibility to perform many different tests on your data

   - Imagine that you plan to conduct 85 contrasts on your data.
     - A Type 1 error is the probability of rejecting the null hypothesis when the null hypothesis is true.
     - If we set α=.05, then if the null hypothesis is true, just by chance alone we will commit 4-5 Type 1 errors

       $$.05 * 85 = 4.25$$

     - However, we have no way of knowing which 4-5 test results are errors!

   - This number of Type 1 errors seems wrong to many people.

     Perhaps we should not think of controlling α at the level of the individual contrast, but at the level of the entire experiment.

2. Types of Type 1 Error rates

   - As a discipline, we have decided that it is very important to maintain the probability of a Type 1 error at .05 or smaller

   - Per-comparison (PC) error rate
     - The probability of committing a Type 1 error for a single contrast

$$\alpha_{PC} = \frac{\textit{The number of contrasts declared falsely significant}}{\textit{number of contrasts}}$$

- Family-wise (FW) error rate
  - Consider a two-way ANOVA with Factor A and Factor B. You can conduct a set of contrasts on the Factor A cell means and on the Factor B cell means.
  - We consider the set of contrasts on Factor A as one family of contrasts, and the set of contrasts on Factor B as a second family of contrasts.
  - The family-wise (FW) error rate is the probability of committing a Type 1 error for an entire family of contrasts

$$\alpha_{FW} = \frac{\textit{The number of families with at least one contrast declared falsely significant}}{\textit{number of families}}$$

- Experiment-wise (EW) error rate
  - The probability of committing at least one Type 1 error over an entire experiment

$$\alpha_{EW} = \frac{\textit{The number of experiments with at least one contrast declared falsely significant}}{\textit{number of experiments}}$$

  - For one-way ANOVA designs, there is only one family and so the α_{EW} equals α_{FW}

- Which α should we be concerned about?
  - One convention is to use the same α for a family of contrasts as was used to test the omnibus null hypothesis for that family
    - That is, if you use $\alpha = .05$ for the omnibus test, then the probability of making a type one error on the entire set of contrasts on that factor should be $\alpha_{FW} = .05$
    - In other words, this convention is to control the <u>family-wise error rate</u>

- A second convention is to control α_{EW} at 5%
  - The experiment seems to be a better conceptual unit than the family
  - Most statisticians agree that we should be more concerned about the experiment-wise error rate

- If you control α_{FW}=.05, then for a two factor ANOVA there are three families of contrasts you have
  - Tests on Factor A α_{FW}=.05
  - Tests on Factor B α_{FW}=.05
  - Tests on the interaction between A and B α_{FW}=.05

  Then $\alpha_{EW} > .05$

- In a sense isn't this all pretty silly?
  - Should journal editors require article-wise error rates?
  - Maybe have journal volume-wise error rates?
  - Perhaps we can have department-wise error rates?
  - The super conscientious researcher might consider a career-wise error rate

- Some additional terminology:
  - The entire reason for monitoring the Type I error rate is to make sure α_{EW} (or α_{FW}) are equal to .05

  - If $\alpha_{EW} < .05$, then the statistical test is said to be **conservative**
    - We will make fewer Type I errors than we are "allowed"
    - But this will also result in a decrease in power

  - If $\alpha_{EW} > .05$, then the statistical test is said to be **liberal**
    - We will make more Type I errors than we are "allowed"
    - The whole point of monitoring the error rate is to avoid this case

- How much of a difference does it make?
  - In general, we do not know when we make a Type I error. Thus, the following are hypothetical in which we know when a Type I error has been made.

  i. An example
  - Suppose that a one-way ANOVA is replicated 1000 times. In each experiment 10 contrasts are tested.
  - Suppose a total of 90 Type 1 Errors are made, and at least one Type 1 Error is made in 70 of the experiments

$$\alpha_{PC} = \frac{90}{10 * 1000} = .009$$

$$\alpha_{FW} = \alpha_{EW} = \frac{70}{1000} = .07$$

  ii. Calculating error rates
  - How do we calculate a family-wise error rate?

$$\alpha_{FW} = \frac{\textit{The number of families with at least one contrast declared falsely significant}}{\textit{number of families}}$$

  - We are interested in the probability of <u>at least one</u> Type 1 Error over a set of contrasts.

  - Let's conduct c independent contrasts, each with α_{PC} =.05

$$\begin{aligned}
\alpha_{FW} &= \text{P(at least one false test result in all c tests)} \\
&= 1 - \text{P(no false test results in all c tests)} \\
&= 1 - \begin{vmatrix} \text{P (no false result in test 1)} * \\ \text{P (no false result in test 2)} * \\ \ldots \quad * \\ \text{P (no false result in test c)} \end{vmatrix} \\
&= 1 - \left[(1-\alpha_{\text{PC}}) * (1-\alpha_{\text{PC}}) * \ldots * (1-\alpha_{\text{PC}})\right] \\
&= 1 - (1-\alpha_{\text{PC}})^{c}
\end{aligned}$$

- In our case we have α_{PC}=.05

$$\alpha_{FW} = 1 - (1 - .05)^c$$

- The family-wise error rate will depend on the number of contrasts we run in that family

# of tests	FW Type 1 error rate
2	.098
3	.143
5	.226
10	.401
15	.537
20	.642

- Even a relatively small number of contrasts can result in a very inflated family-wise error rate!

It turns out that it is even more complicated than just controlling the family-wise or experiment-wise error rate

3. Planned contrasts vs. Post hoc contrasts
   - Planned contrast
     - A contrast that you decided to test prior to an examination of the data
     - Sometimes called a priori tests
     - These comparisons are theory driven
     - Part of a strategy of confirmatory data analysis

   - Post-hoc test
     - A contrast that you decide to test only after observing all or part of the data
     - Sometimes called a posteriori tests
     - These comparisons are data driven
     - Part of an exploratory data analysis strategy

   - Is there really any difference between a planned contrast and a post-hoc contrast?
     - An investigator runs an experiment with four levels: A, B, C, D.
     - Experimenter 1
       - Before the study is run, Experimenter 1 has the following hypothesis:

$$\frac{\mu_1 + \mu_2}{2} = \frac{\mu_3 + \mu_4}{2}$$

     - Experimenter 2
       - Experimenter 2 has no real hypotheses (?!?)
       - After running the study, the following data are observed:

Group			
A	B	C	D
2.0	1.5	5.0	6.0

       - Now Experimenter 2 decides to test the following hypothesis:

$$\frac{\mu_1 + \mu_2}{2} = \frac{\mu_3 + \mu_4}{2}$$

- Are Experimenter 1 and Experimenter 2 testing the same hypothesis?

- Imagine that a different set of data was observed:

Group			
A	B	C	D
2.0	5.0	1.5	6.0

- How would that change the analyses conducted by Experimenter 1 and Experimenter 2?
  - Experimenter 1 had an a priori hypothesis, and would still want to test this hypothesis:

$$\frac{\mu_1 + \mu_2}{2} = \frac{\mu_3 + \mu_4}{2}$$

  - But Experimenter 2 had no a priori hypothesis. He/she will look at the new data and decide to test

$$\frac{\mu_1 + \mu_3}{2} = \frac{\mu_2 + \mu_4}{2}$$

  - Experimenter 2's choice of contrast is determined by the ordering of the cell means. Thus, the hypothesis that Experimenter 2 is actually testing is:

$$\frac{\mu_{(\min)} + \mu_{(\min+1)}}{2} = \frac{\mu_{(\max)} + \mu_{(\max-1)}}{2}$$

- Imagine that the null hypothesis is true and that all differences in means are due to chance.

  - Experimenter 1's true α rate will be .05
  - Experimenter 2's comparisons capitalize on the chance variation in the data. As a result, the probability of committing a Type 1 error rate will be much greater than .05

  - Put another way, when we conduct our statistical tests, we construct hypothetical sampling distributions. The sampling distributions for the following two hypotheses will be very different:

$$\frac{\mu_1 + \mu_3}{2} = \frac{\mu_2 + \mu_4}{2} \qquad \frac{\mu_{(\min)} + \mu_{(\min+1)}}{2} = \frac{\mu_{(\max)} + \mu_{(\max-1)}}{2}$$

4. Planned contrasts

   Let us first consider the situation where you have a set of planned contrasts

   - We do not need to worry about the role of chance in determining our contrasts
   - We may need to worry about inflated experiment-wise error rates

   - Bonferroni Correction (Sometimes called Dunn's test [1961])
     - If we are going to perform *c* orthogonal contrasts, then we simply divide our α_{EW} into *c* parts

       $$\alpha_{PC} = \frac{\alpha_{EW}}{c}$$

     - In most cases, we will want to set $\alpha_{EW} = .05$

       $$\alpha_{PC} = \frac{.05}{c}$$

     - The details of how to apply this correction will follow shortly

   - Dunn/Sidák Correction (1967)
     - The Bonferroni turns out to be slightly too conservative
     - A second approach is based on our previous calculations of α_{FW}

       $$\alpha_{EW} = 1 - (1 - \alpha_{PC})^c$$

     - We would like to control the α_{FW}, so let's solve for α_{PC}

       $$\alpha_{EW} = 1 - (1 - \alpha_{PC})^c$$
       $$(1 - \alpha_{PC})^c = 1 - \alpha_{EW}$$
       $$1 - \alpha_{PC} = \sqrt[c]{1 - \alpha_{EW}}$$
       $$\alpha_{PC} = 1 - \sqrt[c]{1 - \alpha_{EW}}$$

     - This can also be written as $\alpha_{PC} = 1 - (1 - \alpha_{EW})^{\frac{1}{c}}$

     - In most cases, we will want to set $\alpha_{FW} = .05$

       $$\alpha_{PC} = 1 - (.95)^{\frac{1}{c}}$$

- Using either the Dunn/Sidák or Bonferroni correction, what should be used as α_{PC} to keep $\alpha_{EW} = .05$?

# of tests	Dunn/Sidák $1-(1-\alpha)^{\frac{1}{c}}$	Bonferroni α/c
3	.0170	.0167
5	.0102	.0100
10	.0051	.0050
15	.0034	.0033
20	.0026	.0025
25	.0021	.0020
50	.0010	.0010
100	.0005	.0005

  - To maintain $\alpha_{EW} = .05$, you need to use this modified critical *p*-value

  - If you have 10 contrasts of interest
    - Test the contrasts using ONEWAY or UNIANOVA to obtain the exact $p_{observed}$
    - Use the corrected p_{crit} to determine significance
      For $c = 10$, use $p_{crit} = .005$
    - If $p_{observed} < p_{crit}$ then report the test significant at $\alpha_{EW} = .05$

- Miscellaneous notes on both procedures
  - All of our calculations have been based on independent (orthogonal) contrasts. These adjustments can be used for non-orthogonal contrasts and they will be conservative (they will overcorrect)

  - If you have unequal variances in your groups, you can use the unequal variance test for contrasts and then apply these corrections.

- For these procedures, we divide $\alpha_{EW} = .05$ into *c* equal parts
  - Statistically, there is no reason why you have to use equal divisions of α_{EW}. All that is required is that the three parts sum to $\alpha_{EW} = .05$

  - If you are testing 3 contrasts with $\alpha_{EW} = .05$, then you could use the following p_{crit}

$$\begin{array}{ll} \psi_1 & \alpha_{PC} = .03 \\ \psi_2 & \alpha_{PC} = .01 \\ \psi_3 & \alpha_{PC} = .01 \end{array}$$

  - If ψ_1 is more important than ψ_2 and ψ_3, then this unequal splitting of α_{EW} gives you more power to detect a difference for ψ_1
  - The catch is that you must decide how to divide $\alpha_{EW} = .05$ before looking at the data
  - Although this unequal splitting of α_{EW} is statistically legal, there is not a chance you could get this method by a journal editor

- Arguments against the need to correct for a <u>small number</u> of planned contrasts
  - The omnibus F-test is equal to the average of *a-1* orthogonal contrasts (where a is the number of groups)
  - Some behavioral statisticians have argued that if you have *a-1* orthogonal contrasts, there is no need to correct the α-level (and some have extended this argument to apply to *a-1* non-orthogonal contrasts)

  - The (ignored) multi-factor ANOVA problem: For a two factor ANOVA there are three families of contrasts you have
    - Tests on Factor A $\alpha_{FW} = .05$
    - Tests on Factor B $\alpha_{FW} = .05$
    - Tests on the interaction between A and B $\alpha_{FW} = .05$

    As a result, $\alpha_{EW} > .05$

- The story of Bonehead and Bright (with thanks to Rich Gonzalez)
  - Two researchers want to examine 4 treatments. Specifically, they want to compare Treatment A to Treatment B, and Treatment C to Treatment D

  - Bonehead designs two studies

    Study 1: Treatment A vs. Treatment B

    Study 2: Treatment C vs. Treatment D

    Bonehead conducts t-tests to compare the treatments in each study

  - Bright designs one study so that he/she can compare all four treatments. He/she tests the key hypotheses using contrasts

	Treatment			
	A	B	C	D
ψ_1	1	-1		
ψ_2			1	-1

    But now reviewers scream that Bright has conducted two contrasts so the overall Type I error rate will be greater than .05. They demand that Bright use a Bonferroni correction.

    Why should Bright be penalized for designing a better study?

- Thus, if you conduct a small number of planned contrasts (no more than a-1), I believe that no p-value correction is necessary.

  - This issue is still relatively controversial (After all, the experiment-wise Type 1 Error rate will be greater than .05 if you use no correction).
  - However, I believe that you should be rewarded for having strong hypotheses and for planning to conduct tests of those hypotheses.
  - Do not be surprised if you encounter a reviewer who requests a Bonferroni correction for any planned contrast. Be prepared to argue why no correction is necessary

  - If you have more than a-1 planned contrasts, then most people will think you are fishing and will require you to use a Bonferroni correction

- Planned contrasts in SPSS
  - SPSS has a "Bonferroni" and a "Sidak" option under post-hoc tests (even though these tests are not post-hoc tests!)
  - These procedures are only good for if you plan to conduct all possible pair-wise comparisons.
  - Your theory should allow you to make more specific hypotheses than all pairwise comparisons. Hence, SPSS will be of little help.
  - In general, we must correct the p-values by hand.

- Example: The effect of strategy training on memory

  Independent samples of six-year-olds and eight-year-olds are obtained. Half of the children in each group are randomly assigned to a memory-strategy training condition while the other half serve as a control.

  In advance the experimenter wants to test the following contrasts:

	Six-year-olds		Eight-year-olds	
	Training	Control	Training	Control
	μ_{6T}	μ_{6C}	μ_{8T}	μ_{8C}
ψ_1	1	-1		
ψ_2			1	-1
ψ_3	.5	-.5	.5	-.5

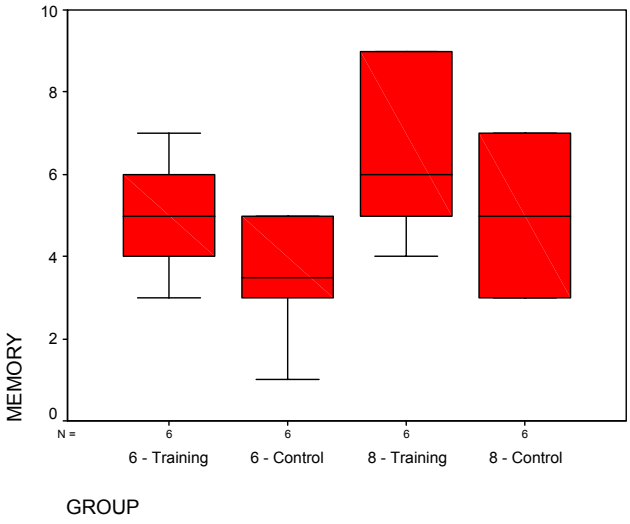
  - Approach 1
    - Conduct the omnibus F-test
    - If non-significant, then stop
    - If significant, then you can test the three hypothesized contrasts
    - There is no justification for this approach

  - Approach 2 (Andy's preference)
    - Skip the omnibus F-test
    - Directly test the planned contrasts, using $\alpha_{PC} = .05$
    - You are rewarded for planning a small number of contrasts and do not have to correct the *p*-values

  - Approach 3
    - Skip the omnibus F-test
    - Directly test the planned contrasts using a Bonferroni correction
    - This approach keeps $\alpha_{EW} \leq .05$

- Now suppose that we run the study and obtain the following data:

Six-year-olds		Eight-year-olds	
Training	Control	Training	Control
μ_{6T}	μ_{6C}	μ_{8T}	μ_{8C}
6	5	6	3
5	3	9	7
7	1	9	6
5	5	4	3
3	3	5	4
4	4	6	7
5.0	**3.5**	**6.5**	**5.0**



  - Approach 1: Conduct the omnibus F-test

```
ONEWAY memory BY cond
  /STAT DESC.
```

ANOVA

MEMORY

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	27.000	3	9.000	2.951	.057
Within Groups	61.000	20	3.050		
Total	88.000	23			

Using this approach, we would stop and never get to test our hypothesis!

- Approach 2: Go directly to uncorrected planned contrasts

```
ONEWAY memory BY group
  /CONT 1 -1 0 0
  /CONT 0 0 1 -1
  /CONT .5 -.5 .5 -.5.
```

Contrast Tests

		Contrast	Value of Contrast	Std. Error	t	df	Sig. (2-tailed)
MEMORY	Assume equal variances	1	1.5000	1.00830	1.488	20	.152
		2	1.5000	1.00830	1.488	20	.152
		3	1.5000	.71297	2.104	20	.048

We find a significant result for contrast 3: the comparison of control vs. treatments across both age groups.

- Approach 3: Go directly to corrected planned contrasts

```
ONEWAY memory BY group
  /POSTHOC = BONFERRONI SIDAK ALPHA(.05).
```

Multiple Comparisons

Dependent Variable: MEMORY

	(I) GROUP	(J) GROUP	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound
Bonferroni	6 - Training	6 - Control	1.5000	1.00830	.915	-1.4514	4.4514
		8 - Training	-1.5000	1.00830	.915	-4.4514	1.4514
		8 - Control	.0000	1.00830	1.000	-2.9514	2.9514
	6 - Control	6 - Training	-1.5000	1.00830	.915	-4.4514	1.4514
		8 - Training	-3.0000*	1.00830	.045	-5.9514	-.0486
		8 - Control	-1.5000	1.00830	.915	-4.4514	1.4514
	8 - Training	6 - Training	1.5000	1.00830	.915	-1.4514	4.4514
		6 - Control	3.0000*	1.00830	.045	.0486	5.9514
		8 - Control	1.5000	1.00830	.915	-1.4514	4.4514
	8 - Control	6 - Training	.0000	1.00830	1.000	-2.9514	2.9514
		6 - Control	1.5000	1.00830	.915	-1.4514	4.4514
		8 - Training	-1.5000	1.00830	.915	-4.4514	1.4514
Sidak	6 - Training	6 - Control	1.5000	1.00830	.629	-1.4418	4.4418
		8 - Training	-1.5000	1.00830	.629	-4.4418	1.4418
		8 - Control	.0000	1.00830	1.000	-2.9418	2.9418
	6 - Control	6 - Training	-1.5000	1.00830	.629	-4.4418	1.4418
		8 - Training	-3.0000*	1.00830	.044	-5.9418	-.0582
		8 - Control	-1.5000	1.00830	.629	-4.4418	1.4418
	8 - Training	6 - Training	1.5000	1.00830	.629	-1.4418	4.4418
		6 - Control	3.0000*	1.00830	.044	.0582	5.9418
		8 - Control	1.5000	1.00830	.629	-1.4418	4.4418
	8 - Control	6 - Training	.0000	1.00830	1.000	-2.9418	2.9418
		6 - Control	1.5000	1.00830	.629	-1.4418	4.4418
		8 - Training	-1.5000	1.00830	.629	-4.4418	1.4418

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

However, SPSS uses c = {all possible pairwise contrasts!}
Because SPSS is of no help, we must resort to hand calculation

- We are conducting 3 contrasts, so we use $c = 3$

	Dunn/Sidák	Bonferroni
# of tests	$1-(1-.05)^{\frac{1}{c}}$	.05/c
3	.0170	.0167

- Compare the observed p-value to these adjusted p-values. We report the test as being significant or not at the $\alpha_{EW} = .05$ level.
- For post p-value correction we will not be able to compute exact, adjusted p-values. We will only be able to report if they tests are significant or not.

- However, for with the Bonferroni and Dunn/Sidák procedures, adjusted p-values may be estimated:

	Estimated experiment-wise adjusted p-value	
	Bonferroni	Dunn/ Sidák
	$p_{unadj} = \frac{p_{adj}}{3}$	$p_{unadj} = 1-\left(1-p_{adj}\right)^{\frac{1}{c}}$
$\hat{\psi}_1 : t(20) = 1.488, p = .152$	$.152 = \frac{p_{adj}}{3}$	$.152 = 1-\left(1-p_{adj}\right)^{\frac{1}{3}}$
$p_{obs} = .152 > .0167 = p_{crit}$	$p_{adj} = .456$	$p_{adj} = .390$
$\hat{\psi}_2 : t(20) = 1.488, p = .152$		
$p_{obs} = .152 > .0167 = p_{crit}$	$p_{adj} = .456$	$p_{adj} = .390$
$\hat{\psi}_3 : t(20) = 2.104, p = .048$	$.048 = \frac{p_{adj}}{3}$	$.048 = 1-\left(1-p_{adj}\right)^{\frac{1}{3}}$
$p_{obs} = .048 > .0167 = p_{crit}$	$p_{adj} = .144$	$p_{adj} = .137$

- We report that with a Bonferroni correction, $\hat{\psi}_1$ $\hat{\psi}_2$, and $\hat{\psi}_3$ are not statistically significant, *all p's* $> .13$

- Bonferroni and Dunn/Sidák 95% confidence intervals
  - We need to compute an adjusted critical value

$$\hat{\psi} \pm \left(t_{criticaladjusted}(dfw) * \sqrt{MSW \sum \frac{c_i^2}{n_i}} \right)$$

  - Bonferroni

$$t_{critical\ adjusted} = t\left(\alpha = \frac{.05}{c}, df = dfw = N - a \right)$$
$$t_{critical\ adjusted} = t(\alpha = .0167, df = 20) = 2.613$$

  - Dunn/Sidák

$$t_{critical\ adjusted} = t\left(\alpha = [1 - (1 - .05)^{\frac{1}{c}}], df = dfw = N - a \right)$$
$$t_{critical\ adjusted} = t(\alpha = .0170, df = 20) = 2.603$$

(These adjusted t-values can be obtained from EXCEL)

	Bonferroni	Dunn/Sidák
$\hat{\psi}_1$	$\hat{\psi} \pm \left(t_{criticaladjusted}(dfw) * \sqrt{MSW \sum \frac{c_i^2}{n_i}} \right)$	$\hat{\psi} \pm \left(t_{criticaladjusted}(dfw) * \sqrt{MSW \sum \frac{c_i^2}{n_i}} \right)$
	$1.50 \pm \left(2.613 * \sqrt{3.05\left(\frac{1}{6} + \frac{1}{6}\right)} \right)$	$1.50 \pm \left(2.603 * \sqrt{3.05\left(\frac{1}{6} + \frac{1}{6}\right)} \right)$
	1.50 ± 2.635	1.50 ± 2.625
	$(-1.14, 4.14)$	$(-1.12, 4.12)$
$\hat{\psi}_3$	$1.50 \pm \left(2.613 * \sqrt{3.05\left(\frac{.25}{6} + \frac{.25}{6} + \frac{.25}{6} + \frac{.25}{6}\right)} \right)$	$1.50 \pm \left(2.603 * \sqrt{3.05\left(\frac{.25}{6} + \frac{.25}{6} + \frac{.25}{6} + \frac{.25}{6}\right)} \right)$
	1.50 ± 1.860	1.50 ± 1.856
	$(-0.36, 3.36)$	$(-0.36, 3.36)$

5. Pairwise Post-hoc tests

- See Kirk (1995) for a nice review of all of these procedures and more!
  [He covers 22 procedures!]

- The Bonferroni and Dunn/Sidák corrections are not post-hoc tests
  - They control the multiple comparison problem
  - They do not directly address the data-driven comparison problem

  - A modification of these methods can be used for *pairwise* post-hoc tests
    - You must let C = the total number of possible pairwise comparisons
    - If you have a groups, then
    $$C = \frac{a(a-1)}{2}$$
    - If any of the following conditions apply, you must use $C = \frac{a(a-1)}{2}$
      i. All pairwise comparisons are to be tested
      ii. The original intent was to test all pairwise comparisons, but after looking at the data, fewer comparisons are actually tested
      iii. The original intent was to compare a subset of all possible pairwise comparisons, but after looking at the data, one or more additional pairwise comparisons are also to be tested

  - So long as $C = \frac{a(a-1)}{2}$, the Bonferroni and Dunn/Sidák corrections can be used in a post-hoc manner.
    - Technically, these methods are not post-hoc corrections. In this case they *can* be used for all pairwise comparisons.
    - But the Bonferroni and Dunn/Sidák corrections are too conservative and less powerful than other techniques that have been developed specifically to test all pairwise comparisons

- The stop-light approach to post-hoc tests
  - The traditional approach to post-hoc tests is to use the omnibus F-test as a post-hoc traffic signal
    - If the omnibus test is significant, then you have a green light to conduct follow-up post-hoc tests
    - If the omnibus test is not significant, then you have a red light. STOP! You are not allowed to proceed with post-hoc tests

  - This line of thought applies to some post-hoc tests, but not all of them. Some researchers (and statisticians) have mistakenly generalized this traffic signal approach to all post hoc tests.

i. Fisher's LSD [Least Significant Difference] (1935)
   - One of the first attempts to solve the multiple comparison problem
   - Fisher developed an approximate solution to this problem based on the omnibus traffic signal mentality
   - He reasoned that if the omnibus test is significant at the .05 level, then α_{EW} is preserved at the .05 level. Thus a significant omnibus test allows you to perform all follow-up contrasts using uncorrected t-tests
   - This test is sometimes called the "protected t-test"
     - Step 1: Perform omnibus F-test.
       If significant, then proceed to Step 2.
       If not significant, then stop.
     - Step 2: Compare all means using uncorrected pairwise contrasts..

   - Unfortunately, this test does not control α_{EW}, and tends to be too liberal. The LSD procedure keeps $\alpha_{EW} \leq .05$ only if there are three or fewer groups in the design. For larger designs, the LSD should be avoided.

- An example of why Fisher's LSD is not valid

- You **should not** use this test and should be very suspicious of anyone how does use it (unless there are only three groups in your design). It is a historical relic that has been replaced by more appropriate tests

ii. Tukey's HSD (Honestly Significant Difference) [1953]
   - I'll present the Tukey-Kramer (1956) test for use when the sample sizes are unequal
   - Tukey realized the problem of post-hoc tests had to do with comparisons being determined by the rank ordering of the cell means

   - For a pairwise contrast, we have the following formula:

$$t_{observed} = \frac{\bar{X}_i - \bar{X}_j}{\sqrt{MSW\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}}$$

   - But in the case of a post hoc comparison, means are selected for comparison based on their ranks

$$t_{Pairwise\ Maximum} = \frac{\bar{X}_{MAX} - \bar{X}_{MIN}}{\sqrt{MSW\left(\frac{1}{n_{MAX}} + \frac{1}{n_{MIN}}\right)}}$$

     - In any dataset, the largest *t* we can observe is given by $t_{Pairwise\ Maximum}$.
     - If we could determine some critical value (CRIT) such that under the null hypothesis, $t_{Pairwise\ Maximum}$ > CRIT only 5% of the time, then because no contrast can be larger than $t_{Pairwise\ Maximum}$, we will have found a CRIT that will keep $\alpha_{EW} = .05$
     - Tukey's insight was to determine a sampling distribution related to $t_{Pairwise\ Maximum}$ called the studentized range, *q*

$$q = \sqrt{2}t$$

- Critical values for the studentized range are given the Appendix of most advanced ANOVA books
- Using the studentized range, we can be sure that under the null hypothesis the largest comparison we could observe will be significant 5% of the time
- For pairwise comparisons less than the Min vs. the Max, the probability of a Type 1 error is less than .05.
- In this manner we can control α_{EW} for the set of all possible pairwise contrasts.
- Tukey's HSD keeps $\alpha_{EW} = .05$ for the largest pairwise contrast, and is conservative for all other comparisons.

- Implementing Tukey's HSD procedure by hand

  - Calculate $t_{observed}$ for any/all pairwise comparisons of interest
  - Look up the critical value $q(1-\alpha, a, \nu)$

    Where α = Familywise error rate
    a = Number of groups
    ν = DFw = $N-a$

  - Compare $t_{observed}$ to $\frac{q_{crit}}{\sqrt{2}}$

  - Note: Because EXCEL (and most other programs) do not compute studentized range p-values, we can not compute exact Tukey adjusted p-values.

- Implementing Tukey's HSD in SPSS

```
ONEWAY dv BY iv
 /POSTHOC = TUKEY ALPHA(.05).
```

  The p-values are exact Tukey-adjusted p-values

- What to do when the population variances are unequal
  Dunnett's T3 procedure (1980)
  - Compute the unequal variances test (with its adjusted degress of freedom) for all pairwise contrasts
  - Estimate the Dunnett's T3 critical value, $q(1-\alpha, a, \nu)$, for the unequal variances test result

    Where α = Familywise error rate
    a = Number of groups
    ν = the variance-adjusted df

  - Compare $t_{observed(adjusted)}$ to $\frac{q_{crit}}{\sqrt{2}}$

  - SPSS's Dunnett's T3 procedure is not much help because it does not output the adjusted degrees of freedom, so you do not have enough information to report the adjusted tests!

- An example of Tukey's HSD

  - Suppose we would like to examine all pairwise comparisons in the memory-training example.

- Step 1: Calculate $t_{observed}$ for any/all pairwise comparisons of interest
  For simplicity, I will select two pairwise contrasts, but we are allowed to conduct all pairwise contrasts.

$\psi_1 : \mu_{8T} - \mu_{8C}$ $\qquad t_{observed} = \dfrac{\bar{X}_i - \bar{X}_j}{\sqrt{MSW\left(\dfrac{1}{n_i} + \dfrac{1}{n_j}\right)}} = \dfrac{6.5 - 5.0}{\sqrt{3.05\left(\dfrac{1}{6} + \dfrac{1}{6}\right)}} = 1.50$

$\psi_2 : \mu_{8T} - \mu_{6C}$ $\qquad t_{observed} = \dfrac{\bar{X}_i - \bar{X}_j}{\sqrt{MSW\left(\dfrac{1}{n_i} + \dfrac{1}{n_j}\right)}} = \dfrac{6.5 - 3.5}{\sqrt{3.05\left(\dfrac{1}{6} + \dfrac{1}{6}\right)}} = 2.98$

- Step 2: Look up the critical value $q(1-\alpha, a, \nu)$
  Where α = Experiment-wise error rate
  a = Number of groups
  ν = DFw = N-a

$q(1-\alpha, a, \nu) = q(.95, 4, 20) = 3.96$

$\dfrac{q_{crit}}{\sqrt{2}} = \dfrac{3.96}{\sqrt{2}} = 2.80$

- Step 3: Compare $t_{observed}$ to $\dfrac{q_{crit}}{\sqrt{2}}$

$\psi_1 : \mu_{8T} - \mu_{8C}$ $\qquad t_{obs} = 1.50 < 2.80 = \dfrac{q_{crit}}{\sqrt{2}}$ $\qquad$ Fail to Reject H_0

$\psi_2 : \mu_{8T} - \mu_{6C}$ $\qquad t_{obs} = 2.98 > 2.80 = \dfrac{q_{crit}}{\sqrt{2}}$ $\qquad$ Reject H_0

- Tukey 95% confidence intervals
  - We need to compute an adjusted critical value

$$\hat{\psi} \pm \left(t_{criticaladjusted} * \sqrt{MSW\left(\frac{1}{n_i} + \frac{1}{n_j}\right)} \right)$$

$$t_{critical\ adjusted} = \frac{q(\alpha, a, dfw)}{\sqrt{2}}$$

$$t_{critical\ adjusted} = \frac{q(.05, 4, 20)}{\sqrt{2}} = \frac{3.96}{\sqrt{2}} = 2.80$$

$\psi_1 : \mu_{8T} - \mu_{8C}$

$$\hat{\psi} \pm \left(t_{criticaladjusted} * \sqrt{MSW\left(\frac{1}{n_i} + \frac{1}{n_j}\right)} \right)$$

$$1.50 \pm \left(2.80 * \sqrt{3.05\left(\frac{1}{6} + \frac{1}{6}\right)} \right)$$

$$1.50 \pm 2.823$$

$$(-1.32, 4.32)$$

$\psi_2 : \mu_{8T} - \mu_{6C}$

$$\hat{\psi} \pm \left(t_{criticaladjusted} * \sqrt{MSW\left(\frac{1}{n_i} + \frac{1}{n_j}\right)} \right)$$

$$2.98 \pm \left(2.80 * \sqrt{3.05\left(\frac{1}{6} + \frac{1}{6}\right)} \right)$$

$$3.00 \pm 2.823$$

$$(0.16,\ 5.80)$$

- An example of Tukey's HSD using SPSS

```
ONEWAY memory BY group
  /POSTHOC = TUKEY.
```

Multiple Comparisons

Dependent Variable: MEMORY

Tukey HSD

(I) GROUP	(J) GROUP	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
6 - Training	6 - Control	1.5000	1.00830	.463	-1.3222	4.3222
	8 - Training	-1.5000	1.00830	.463	-4.3222	1.3222
	8 - Control	.0000	1.00830	1.000	-2.8222	2.8222
6 - Control	6 - Training	-1.5000	1.00830	.463	-4.3222	1.3222
	8 - Training	-3.0000*	1.00830	.035	-5.8222	-.1778
	8 - Control	-1.5000	1.00830	.463	-4.3222	1.3222
8 - Training	6 - Training	1.5000	1.00830	.463	-1.3222	4.3222
	6 - Control	3.0000*	1.00830	.035	.1778	5.8222
	8 - Control	1.5000	1.00830	.463	-1.3222	4.3222
8 - Control	6 - Training	.0000	1.00830	1.000	-2.8222	2.8222
	6 - Control	1.5000	1.00830	.463	-1.3222	4.3222
	8 - Training	-1.5000	1.00830	.463	-4.3222	1.3222

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

MEMORY

Tukey HSD[a]

GROUP	N	Subset for alpha = .05	
		1	2
6 - Control	6	3.5000	
6 - Training	6	5.0000	5.0000
8 - Control	6	5.0000	5.0000
8 - Training	6		6.5000
Sig.		.463	.463

Means for groups in homogeneous subsets are displayed.

a. Uses Harmonic Mean Sample Size = 6.000.

- This table gives homogeneous subsets of means using the Tukey procedure

$$\mu_{6C} = \mu_{6T} = \mu_{8C} \qquad \mu_{6T} = \mu_{8C} = \mu_{8T}$$

From these homogeneous sets, we can conclude that

$$\mu_{6C} \neq \mu_{8T}$$

- An example of Dunnett's T3 procedure with unequal variances.
  - Returning to the Bank Starting Salary Data (see 4-45)
  - We determined that these data satisfied the normality assumption, but violated the homogeneity of variances assumption
  - Now we decide to conduct all post-hoc pairwise comparisons

  - The unequal variances version of Tukey's HSD is the Dunnett's T3 test. Let's see what happens when SPSS conducts the T3 procedure

```
ONEWAY salbeg BY jobcat
 /POSTHOC = T3.
```

Multiple Comparisons

Dependent Variable: BEGINNING SALARY

Dunnett T3

(I) EMPLOYMENT CATEGORY	(J) EMPLOYMENT CATEGORY	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
CLERICAL	OFFICE TRAINEE	254.98	156.819	.256	-73.26	583.21
	SECURITY OFFICER	-297.16	294.407	.249	-682.20	87.88
	COLLEGE TRAINEE	-4222.54*	245.408	.000	-5169.67	-3275.41
	EXEMPT EMPLOYEE	-7524.93*	273.079	.000	-9206.23	-5843.62
OFFICE TRAINEE	CLERICAL	-254.98	156.819	.256	-583.21	73.26
	SECURITY OFFICER	-552.14*	304.697	.001	-930.97	-173.31
	COLLEGE TRAINEE	-4477.52*	257.663	.000	-5422.16	-3532.87
	EXEMPT EMPLOYEE	-7779.90*	284.143	.000	-9459.89	-6099.92
SECURITY OFFICER	CLERICAL	297.16	294.407	.249	-87.88	682.20
	OFFICE TRAINEE	552.14*	304.697	.001	173.31	930.97
	COLLEGE TRAINEE	-3925.38*	358.432	.000	-4886.27	-2964.49
	EXEMPT EMPLOYEE	-7227.76*	377.916	.000	-8916.14	-5539.39
COLLEGE TRAINEE	CLERICAL	4222.54*	245.408	.000	3275.41	5169.67
	OFFICE TRAINEE	4477.52*	257.663	.000	3532.87	5422.16
	SECURITY OFFICER	3925.38*	358.432	.000	2964.49	4886.27
	EXEMPT EMPLOYEE	-3302.39*	341.131	.000	-5165.13	-1439.65
EXEMPT EMPLOYEE	CLERICAL	7524.93*	273.079	.000	5843.62	9206.23
	OFFICE TRAINEE	7779.90*	284.143	.000	6099.92	9459.89
	SECURITY OFFICER	7227.76*	377.916	.000	5539.39	8916.14
	COLLEGE TRAINEE	3302.39*	341.131	.000	1439.65	5165.13

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

  - No adjusted Dfs
  - It also appears that the standard errors and confidence intervals are wrong (They assume equal variances)!
  - Because the standard errors are wrong, you cannot compute the adjusted t-value, AND the adjusted degrees of freedom are not reported!

- An alternative is to calculate the T3 procedure by hand

- First calculate all pairwise contrasts (or all pairwise contrasts of interest) using the unequal variance method

```
ONEWAY salbeg BY jobcat
  /CONT -1 1 0 0 0
  /CONT -1 0 1 0 0
  /CONT -1 0 0 1 0
  /CONT -1 0 0 0 1
  /CONT 0 -1 1 0 0
  /CONT 0 -1 0 1 0
  /CONT 0 -1 0 0 1
  /CONT 0 0 -1 1 0
  /CONT 0 0 -1 0 1
  /CONT 0 0 0 -1 1.
```

Contrast Tests

		Contrast	Value of Contrast	Std. Error	t	df	Sig. (2-tailed)
BEGINNING SALARY	Does not assume equal variances	1	-254.98	116.528	-2.188	345.880	.029330
		2	297.16	133.369	2.228	68.848	.029142
		3	4222.54	323.084	13.069	46.034	.000000
		4	7524.93	562.514	13.377	32.443	.000000
		5	552.14	130.812	4.221	62.580	.000080
		6	4477.52	322.037	13.904	45.424	.000000
		7	7779.90	561.913	13.845	32.304	.000000
		8	3925.38	328.506	11.949	48.356	.000000
		9	7227.76	565.645	12.778	33.127	.000000
		10	3302.39	637.613	5.179	49.749	.000004

- Next look up the critical value on the studentized range table, $q(1-\alpha, a, \nu)$

  α = Familywise error rate
  a = Number of groups
  ν = the variance adjusted degrees of freedom

		Critical q	Critical t	Observed t	
Group 1 vs. Group 2:	df = 345.88	3.87	2.73	2.188	ns
Group 1 vs. Group 3:	df = 68.848	3.97	2.81	2.228	ns
Group 1 vs. Group 4:	df = 46.034	4.02	2.84	13.07	*
Group 1 vs. Group 5:	df = 32.443	4.09	2.89	13.38	*

- Compare the Critical t to the Observed t

iii. Dunnett's test (1955)

- Used to compare each of the cell means with a control group mean
- and when the cell sizes are equal
- and when the variances are equal across all groups

- Under these very specific conditions, Dunnett's test controls α_{FW}

- To conduct Dunnett's test
  - Calculate the standard t-test to compare a group mean to the control group
  - Look up Dunnett critical values in a table to determine significance (for example, see Kirk, 1995, Table E.7). Confidence intervals can also be constructed using Dunnett critical values

- Or in SPSS
  - To compare all groups to the last group

```
ONEWAY dv BY iv
  /POSTHOC = DUNNETT.
```

  - To compare all groups to the b^{th} group

```
ONEWAY dv BY iv
  /POSTHOC = DUNNETT (b).
```

  - The p-values are exact Dunnett-adjusted p-values

- It is fine to use this procedure so long as you use it in the appropriate conditions. Modifications have been proposed for use when:
  - The variance of the control group differs from the other variances (Dunnett, 1964)
  - The sample sizes are not all equal (Hochberg & Tamhane, 1987)

iv. Comparing Pairwise Post-Hoc Tests (dfw = 12)

Number of groups	Critical Value			
	Per-Comparison	*Dunnett*	*Tukey*	*Bonferroni*
2	2.18	2.18	2.18	2.18
3	2.18	2.50	2.67	2.78
4	2.18	2.68	2.96	3.14
5	2.18	2.81	3.19	3.43
6	2.18	2.90	3.37	3.65

v. (Student) Newman-Keuls (SNK) test [1939/1952]
- (Unfortunately) a somewhat popular test among psychologists
- Used for pairwise comparisons
- Here's the logic for the SNK test
  - Tukey controls the α_{EW} for the largest possible pairwise comparison
  - But for most pairwise comparisons, Tukey is too conservative
  - Let's develop a correction that is sensitive to different levels of comparison
    - $\Rightarrow$ For means that are farthest apart, use a large critical value
    - $\Rightarrow$ For means that are closer together, use a smaller critical value

- An example: Suppose we have 6 groups
  Let's rank the groups from smallest to largest

  $\mu_{(1)}$ $\mu_{(2)}$ $\mu_{(3)}$ $\mu_{(4)}$ $\mu_{(5)}$ $\mu_{(6)}$

  - If we were to do a Tukey comparison, we would calculate $t_{observed}$ for any/all pairs of interest, and compare the $t_{observed}$ to $\frac{q(\alpha,a,dfw)}{\sqrt{2}}$ where a = number of groups = 6

  - But suppose I want to compare $\mu_{(2)}$ to $\mu_{(5)}$. Let's imagine that $\mu_{(1)}$ and $\mu_{(6)}$ never existed. Then to compare $\mu_{(2)}$ to $\mu_{(5)}$, we would use a critical value of $\frac{q(\alpha,a,dfw)}{\sqrt{2}}$ where $a = 4$.

- For the SNK procedure, the critical value we use is determined NOT by the number of total groups, but by the number of "steps" between the means of interest

Steps between Means	CRIT	Steps between Means	CRIT
1	$\frac{q(\alpha,2,dfw)}{\sqrt{2}}$	4	$\frac{q(\alpha,5,dfw)}{\sqrt{2}}$
2	$\frac{q(\alpha,3,dfw)}{\sqrt{2}}$	5	$\frac{q(\alpha,6,dfw)}{\sqrt{2}}$
3	$\frac{q(\alpha,4,dfw)}{\sqrt{2}}$		

For means 1 step apart: $\frac{q(\alpha,2,dfw)}{\sqrt{2}}$

This is equivalent to CRIT for Fisher's LSD test

For means a-1 steps apart: $\frac{q(\alpha,a,dfw)}{\sqrt{2}}$

This is equivalent to CRIT for Tukey's HSD test

- Although this method sounds appealing, it fails to control α_{EW} (In fact, it is impossible to calculate the exact α_{EW} for SNK). For this reason, most statisticians will not recommend SNK
- It is also not possible to construct SNK confidence intervals

- To implement SNK in SPSS

```
ONEWAY memory BY group
 /POSTHOC = SNK ALPHA(.05).
```

MEMORY

Student-Newman-Keuls[a]

GROUP	N	Subset for alpha = .05	
		1	2
6 - Control	6	3.5000	
6 - Training	6	5.0000	5.0000
8 - Control	6	5.0000	5.0000
8 - Training	6		6.5000
Sig.		.318	.318

Means for groups in homogeneous subsets are displayed.

a. Uses Harmonic Mean Sample Size = 6.000.

- You do not get a significance test for each pair of means, only a list of homogeneous subsets

vi. REGWQ

- The REGWQ test is a modification of the SNK that maintains the experiment-wise error rate at .05 or less.

  - REGWQ was developed by **R**yan (1959) and modified by **E**inot & **G**abriel (1975) and then again by **W**elsh (1977). Its critical value is based on the studentized range, or **Q** distribution.
  - Some authors refer to this procedure as the modified Ryan test.
  - It modifies the α level used to compute critical values for the SNK procedure

For means a-1 steps apart: $\frac{q(\alpha, a, dfw)}{\sqrt{2}}$

For means s steps apart: $\frac{q(\alpha', s+1, dfw)}{\sqrt{2}}$

Where $s = 1, 2, \ldots, a\text{-}1$

$\alpha' = 1 - (1-\alpha)^{s/a}$

  - In general, the REGWQ procedure results in computing a critical q value with a fractional alpha value. Unless you have access to a computer program to calculate exact studentized range values, it is not possible to compute this test by hand.

  - When variances are equal and cell sizes are equal, simulations have shown that the REGWQ procedure keeps $\alpha_{EW} \leq .05$ and is more powerful than Tukey's HSD.

    - (Because we lack the tools to compute this test by hand and check SPSS's calculations, we will not use the REGWQ procedure in this class. I present it because it is a valid option and you may want to use it in your own research)

- To give you an idea how out-of-control post hoc testing can be, Tukey proposed a compromise between the Tukey's HSD and SNK procedures, called Tukey's B

  - Without any theoretical reason, for Tukey's B, you average the critical values from the Tukey's HSD and SNK procedures

$$\text{Critical value for Tukey's B} = \frac{\frac{q(\alpha,a,dfw)}{\sqrt{2}} + \frac{q(\alpha,t,dfw)}{\sqrt{2}}}{2}$$

Where α = Experiment-wise error rate
a = Number of groups
t = Number of steps + 1
ν = DFw = N-a

  - You should **never** use Tukey's B! I only present it as a demonstration of how some post-hoc test have been developed with no theoretical background.

6. Complex Post-hoc tests

i. Scheffé (1953)

- This test is an extension of the Tukey test to all possible comparisons
- The Scheffé test uses a modification of *F* distribution, so we will switch to compute F-tests of contrast

- For the Tukey HSD test, we found the sampling distribution of $F_{Pairwise\ Maximum}$
- But now we want to control the α_{EW} for all possible comparisons
- We need to find the sampling distribution of $F_{Maximum}$, which represents the largest possible F value we could observe for any contrast in the data, either pairwise or complex

- In any data set, we can find a contrast with the sum of squares of the largest possible contrast $\hat{\psi}_{MAX}$ equal to the sum of squares between

$$SS_{\psi MAX} = SSB$$

- Recall that the F-test of a contrast is given by the following formula:

$$F(1, dfw) = \frac{SSC \big/ dfc}{SSW \big/ dfw} = \frac{SSC}{MSW}$$

- Now for $\hat{\psi}_{MAX}$ we have

$$F_{Maximum} = \frac{SSC_{MAX}}{MSW} = \frac{SSB}{MSW}$$

  For any data set, formula gives the F value for the largest possible contrast!

- To find the sampling distribution $F_{Maximum}$, we can use the fact that

$$MSB = \frac{SSB}{a-1}$$

  And then with a bit of algebra . . .

$$SSB = (a-1)MSB$$

$$F_{Maximum} \sim \frac{SSB}{MSW} = \frac{(a-1)MSB}{MSW} = (a-1)F_{\alpha=.05;a-1,N-a}$$

- By using $(a-1)F_{\alpha=.05;a-1,N-a}$ as a critical value for the significance of a contrast, we guarantee α_{EW}=.05, regardless of how many contrasts we test, even after having looked at the data (given that all the assumptions of the F-test have been met)

- There is a direct, one-to-one correspondence between the test of significance of the null hypothesis and the Scheffé test for a contrast
  - If the omnibus F is significant, then there exists at least one contrast that is significant using the Scheffé test
  - If the omnibus F is not significant, then it is impossible to find a significant contrast using the Scheffé test

- Scheffé 95% confidence intervals
  - We need to compute an adjusted critical value

$$\hat{\psi} \pm \left(t_{criticaladjusted} * \sqrt{MSW\left(\sum_{j=1}^{a} \frac{c_j^2}{n_j}\right)} \right)$$

$$t_{critical\ adjusted} = \sqrt{(a-1)F(.05; a-1, Dfw)}$$

- Unfortunately, we cannot use SPSS for Scheffé contrasts
  - The "SCHEFFE" option in SPSS only tests pairwise comparisons
  - The Tukey HSD will always be more powerful than Scheffé to test pairwise comparisons

Number of groups	Critical Value (dfw = 30)		
	Per-Comparison	*Tukey*	*Scheffé*
2	4.17	4.17	4.17
3	4.17	6.09	6.63
4	4.17	7.41	8.77
5	4.17	8.41	10.76
6	4.17	9.25	12.67

  - For complex comparisons, we will have to use hand calculations to determine significance using the Scheffé procedure
  - We can use ONEWAY to find $t_{observed}$ or $F_{observed}$ for the contrast, but we'll have to look-up the critical value and determine significance on our own.

- Using Scheffé for planned contrasts
  - We previously determined that a Bonferroni correction could be used when you have more than a-1 planned contrasts
  - With enough comparisons, the Bonferroni correction is actually more conservative than the Scheffé correction

Number of Planned Comparisons	Critical Value (a=4, dfw = 30)	
	Bonferroni $F(.05/C;1,30)$	*Scheffé* $3F(.05;3,30)$
1	**	**
2	**	**
3	**	**
4	**7.08**	8.76
5	**7.56**	8.76
6	**8.01**	8.76
7	**8.35**	8.76
8	**8.64**	8.76
9	8.94	**8.76**
10	9.18	**8.76**

  - For all planned pairwise comparisons, Tukey is also an option!

- Example #1: The Memory Training experiment
  - The omnibus F-test is not significant
    - No contrasts will be significant using the Scheffé procedure

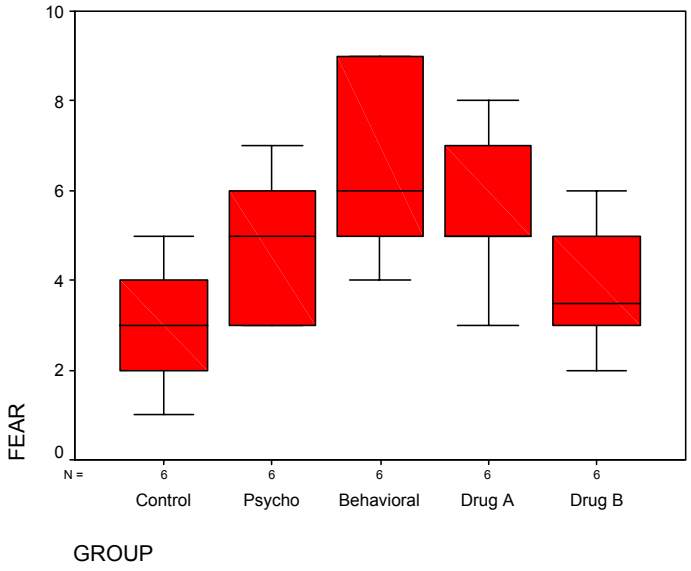
ANOVA

MEMORY

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	27.000	3	9.000	2.951	.057
Within Groups	61.000	20	3.050		
Total	88.000	23			

- Example #2: Treatment of Agoraphobia
  - Twenty-four participants are randomly assigned to one of five conditions for treatment of agoraphobia: a control group, psychodynamic treatment, behavioral treatment, Drug A, or Drug B. The following are post-test scores on a fear scale (lower scores indicate more fear)

Control	Psycho	Behav	Drug A	Drug B
5	3	6	8	6
3	7	9	5	4
2	6	9	7	2
4	5	4	5	3
3	3	5	3	5
1	5	6	5	3
3.0	**4.83**	**6.5**	**5.5**	**3.83**



ANOVA

FEAR

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	45.200	4	11.300	3.998	.012
Within Groups	70.667	25	2.827		
Total	115.867	29			

- After looking at the data, you decide to make the following comparisons:

$\psi_1 = -4\mu_1 + \mu_2 + \mu_3 + \mu_4 + \mu_5$ Control vs. Others

$\psi_2 = \mu_2 + \mu_3 - \mu_4 - \mu_5$ Drug vs. Non-Drug Treatment

$\psi_3 = 2\mu_1 - \mu_3 - \mu_4$ Two best treatments vs. Control

```
ONEWAY fear BY group
  /CONTRAST= -4 1 1 1 1
  /CONTRAST= 0 1 1 -1 -1
  /CONTRAST= -2 0 1 1 0
  /POSTHOC = SCHEFFE ALPHA(.05).
```

- SPSS's useless Scheffé output:

Multiple Comparisons

Dependent Variable: FEAR

Scheffe

(I) GROUP	(J) GROUP	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Control	Psycho	-1.8333	.97068	.483	-5.0578	1.3911
	Behavioral	-3.5000*	.97068	.028	-6.7245	-.2755
	Drug A	-2.5000	.97068	.191	-5.7245	.7245
	Drug B	-.8333	.97068	.944	-4.0578	2.3911
Psycho	Control	1.8333	.97068	.483	-1.3911	5.0578
	Behavioral	-1.6667	.97068	.576	-4.8911	1.5578
	Drug A	-.6667	.97068	.975	-3.8911	2.5578
	Drug B	1.0000	.97068	.897	-2.2245	4.2245
Behavioral	Control	3.5000*	.97068	.028	.2755	6.7245
	Psycho	1.6667	.97068	.576	-1.5578	4.8911
	Drug A	1.0000	.97068	.897	-2.2245	4.2245
	Drug B	2.6667	.97068	.144	-.5578	5.8911
Drug A	Control	2.5000	.97068	.191	-.7245	5.7245
	Psycho	.6667	.97068	.975	-2.5578	3.8911
	Behavioral	-1.0000	.97068	.897	-4.2245	2.2245
	Drug B	1.6667	.97068	.576	-1.5578	4.8911
Drug B	Control	.8333	.97068	.944	-2.3911	4.0578
	Psycho	-1.0000	.97068	.897	-4.2245	2.2245
	Behavioral	-2.6667	.97068	.144	-5.8911	.5578
	Drug A	-1.6667	.97068	.576	-4.8911	1.5578

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

- First compute the value of the contrast and the $F_{observed}$

Contrast Tests

		Contrast	Value of Contrast	Std. Error	t	df	Sig. (2-tailed)
FEAR	Assume equal variances	1	8.6667	3.06956	2.823	25	.009
		2	2.0000	1.37275	1.457	25	.158
		3	6.0000	1.68127	3.569	25	.001
	Does not assume equal variances	1	8.6667	2.71211	3.196	9.158	.011
		2	2.0000	1.42205	1.406	18.691	.176
		3	6.0000	1.60208	3.745	12.875	.002

$\hat{\psi}_1 = 8.667 \qquad F_{observed} = 7.969$
$\hat{\psi}_2 = 2.000 \qquad F_{observed} = 2.123$
$\hat{\psi}_3 = 6.000 \qquad F_{observed} = 12.738$

- Look up $F_{crit} = (a-1)F_{\alpha=.05;a-1,N-a}$

$\alpha = 0.05$
$a = \quad 5$
$N = 30$

$F_{.05}(4,25) = 2.759$
$F_{crit} = 4 * 2.759 = 11.035$

- Compare $F_{observed}$ to F_{crit} in order to determine significance

$\hat{\psi}_1$ Fail to reject null hypothesis
$\hat{\psi}_2$ Fail to reject null hypothesis
$\hat{\psi}_3$ Reject null hypothesis

What are our conclusions?

- Calculate 95% confidence intervals

$$\hat{\psi} \pm \left(t_{criticaladjusted} * \sqrt{MSW\left(\sum_{j=1}^{a} \frac{c_j^2}{n_j}\right)} \right)$$

$$t_{critical\ adjusted} = \sqrt{(a-1)F(.05; a-1, Dfw)}$$
$$t_{critical\ adjusted} = \sqrt{11.035} = 3.32$$

$$8.667 \pm \left(3.32 * \sqrt{2.827\left(\frac{16}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6}\right)} \right)$$
$$8.667 \pm 10.16$$
$$(-1.49, 18.83)$$

$\hat{\psi}_2 = 2.000$

$$2.000 \pm \left(3.32 * \sqrt{2.827\left(\frac{0}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6}\right)} \right)$$
$$2.000 \pm 4.558$$
$$(-2.56, 6.56)$$

$\hat{\psi}_3 = 6.000$

$$6.000 \pm \left(3.32 * \sqrt{2.827\left(\frac{4}{6} + \frac{0}{6} + \frac{1}{6} + \frac{1}{6} + \frac{0}{6}\right)} \right)$$
$$6.000 \pm 5.582$$
$$(0.42, 11.58)$$

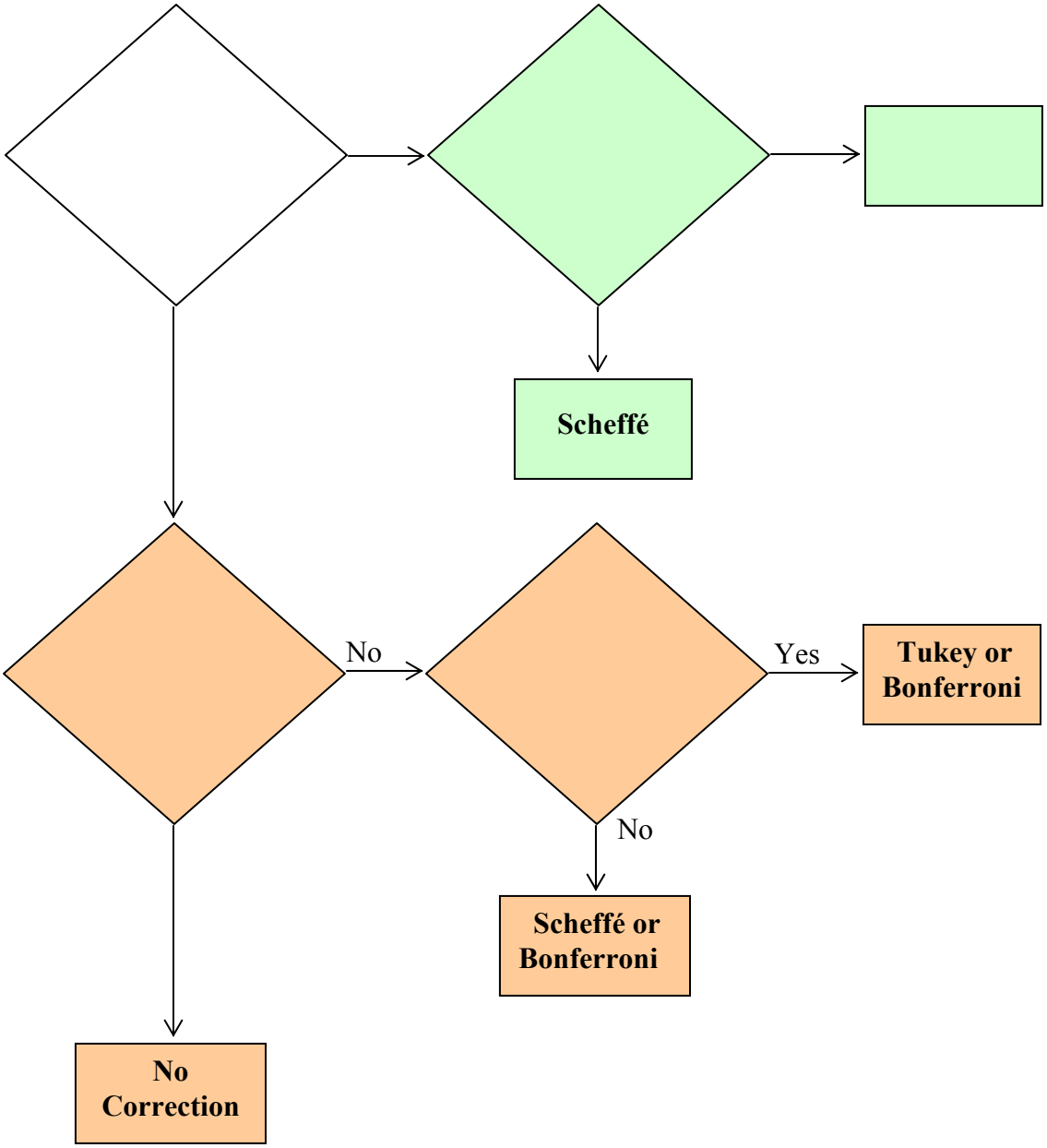
- To get out of the mindset that p-values are everything, do not forget to also report a measure of effect size!

ii. The Brown-Forsythe (1974) test

- The Brown-Forsythe (1974) modification of the Scheffé test that can be used with unequal variances
- Use output from the "variances unequal" line of the ONEWAY command (with an adjusted test statistic and an adjusted degrees of freedom) and proceed with the Scheffé correction in a standard way

7. Conclusions

Which test do I use? A rough guideline



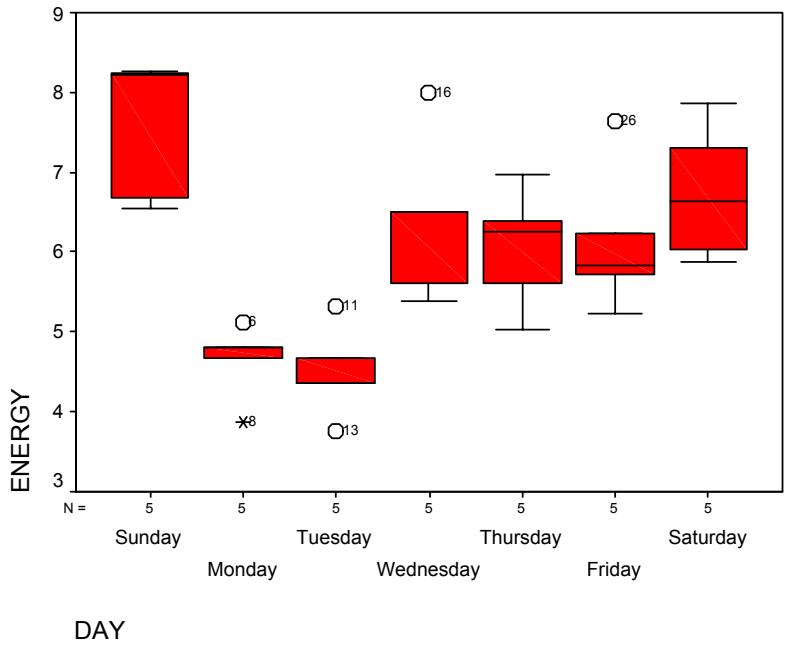
- Modifications of these procedures are necessary when variances are unequal.

8. Examples

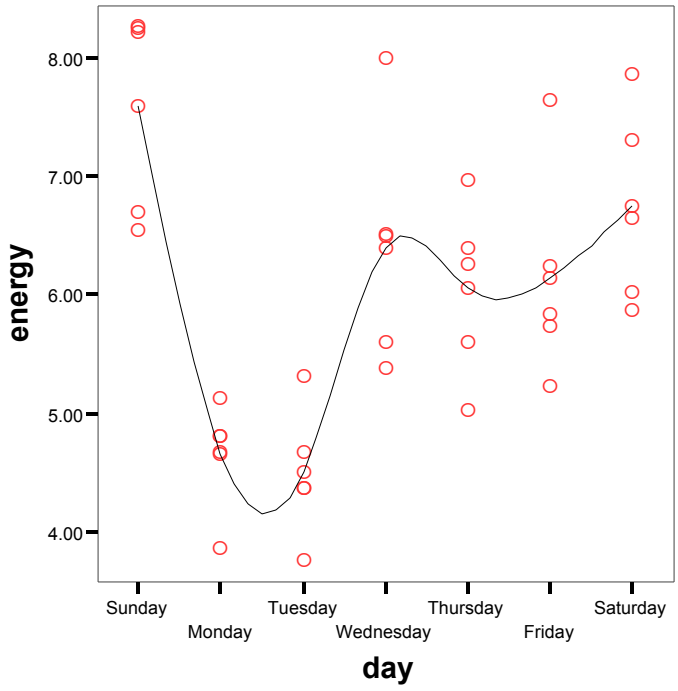
- You want to examine energy use in one-bedroom apartments as a function of the day of the week. You find 35 one-bedroom apartments and measure the energy use in each apartment on a randomly determined day.

Day of the Week						
Sunday	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
8.25	5.12	5.32	8.00	6.97	7.65	7.86
8.26	4.81	4.37	6.50	6.26	5.84	7.31
6.55	3.87	3.76	5.38	5.03	5.23	5.87
8.21	4.81	4.67	6.51	6.40	6.24	6.64
6.69	4.67	4.37	5.60	5.60	5.73	6.03

  - Before the study was run, you decided to look for polynomial trends in energy use.



  - A closer inspection of the assumptions reveals all OK (for *n*'s of 5)



- Because we planned the analysis, we can test the orthogonal polynomial contrasts without correction.

```
ONEWAY energy BY day
 /POLYNOMIAL= 5.
```

(ONEWAY can only test up to the 5th order)

ANOVA

ENERGY

			Sum of Squares	df	Mean Square	F	Sig.
Between Groups	(Combined)		36.633	6	6.106	9.461	.000
	Linear Term	Contrast	.692	1	.692	1.072	.309
	Quadratic	Contrast	12.391	1	12.391	19.201	.000
	Cubic Term	Contrast	12.584	1	12.584	19.501	.000
	4th-order	Contrast	8.713	1	8.713	13.502	.001
	5th-order	Contrast	.059	1	.059	.091	.765
		Deviation	2.195	1	2.195	3.401	.076
Within Groups			18.069	28	.645		
Total			54.702	34			

$\hat{\psi}_{linear}$: $t(28) = 1.03, p = .309, \omega^2 < .001$

$\hat{\psi}_{quadratic}$: $t(28) = 4.38, p < .001, \omega^2 = .212$

$\hat{\psi}_{cubic}$: $t(28) = 4.42, p < .001, \omega^2 = .216$

$\hat{\psi}_{4th}$: $t(28) = 3.67, p = .001, \omega^2 = .146$

$\hat{\psi}_{5th}$: $t(28) = 0.30, p = .765, \omega^2 < .001$

$\hat{\psi}_{6th}$: $t(28) = 1.84, p = .076, \omega^2 = .027$

- If we had done a Bonferroni-type correction,

	Dunn/Sidák	Bonferroni
# of tests	$1-(1-.05)^{\frac{1}{c}}$	.05/c
6	.0085	.0083

  - And our interpretation would not change.

- After looking at the data, we decide to look for polynomial trends only during the weekdays

  - We should not do the following:

```
select if day >1 and day < 7.
ONEWAY energy BY day
  /POLYNOMIAL= 4.
```

  - Why?

```
ONEWAY energy BY day
  /cont = 0 -2 -1 0 1 2 0
  /cont = 0 2 -1 -2 -1 2 0
  /cont = 0 -1 2 0 -2 1 0
  /cont = 0 1 -4 6 -4 1 0.
```

Contrast Tests

		Contrast	Value of Contrast	Std. Error	t	df	Sig. (2-tailed)
ENERGY	Assume equal variances	1	4.5180	1.13606	3.977	28	.000
		2	-1.7580	1.34420	-1.308	28	.202
		3	-1.6260	1.13606	-1.431	28	.163
		4	6.9820	3.00573	2.323	28	.028

- But now, these are post hoc tests and we need to apply the Scheffé correction.
  - Look up $F_{crit} = (a-1)F_{\alpha=.05;a-1,N-a}$

    $\alpha = 0.05$
    $a = \quad 7$
    $N = 35$

    $$F_{crit} = 6 * 2.445 = 14.67$$
    $$t_{crit} = \sqrt{6 * 2.445} = 3.83$$

- Using a Scheffé correction, we find

$\hat{\psi}_{linear}$: $t(28) = 4.00, p < .05$
$\hat{\psi}_{quadratic}$: $t(28) = 1.31, ns$
$\hat{\psi}_{cubic}$: $t(28) = 1.43, ns$
$\hat{\psi}_{4th}$: $t(28) = 2.32, ns$

- We also decide to compare all days to Tuesday:

```
ONEWAY energy BY day
  /posthoc=tukey.
```

Multiple Comparisons

Dependent Variable: ENERGY

Tukey HSD

(I) DAY	(J) DAY	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Tuesday	Sunday	-3.0940	.50806	.000	-4.7056	-1.4824
	Monday	-.1580	.50806	1.000	-1.7696	1.4536
	Wednesday	-1.9000	.50806	.013	-3.5116	-.2884
	Thursday	-1.5540	.50806	.064	-3.1656	.0576
	Friday	-1.6400	.50806	.044	-3.2516	-.0284
	Saturday	-2.2440	.50806	.002	-3.8556	-.6324

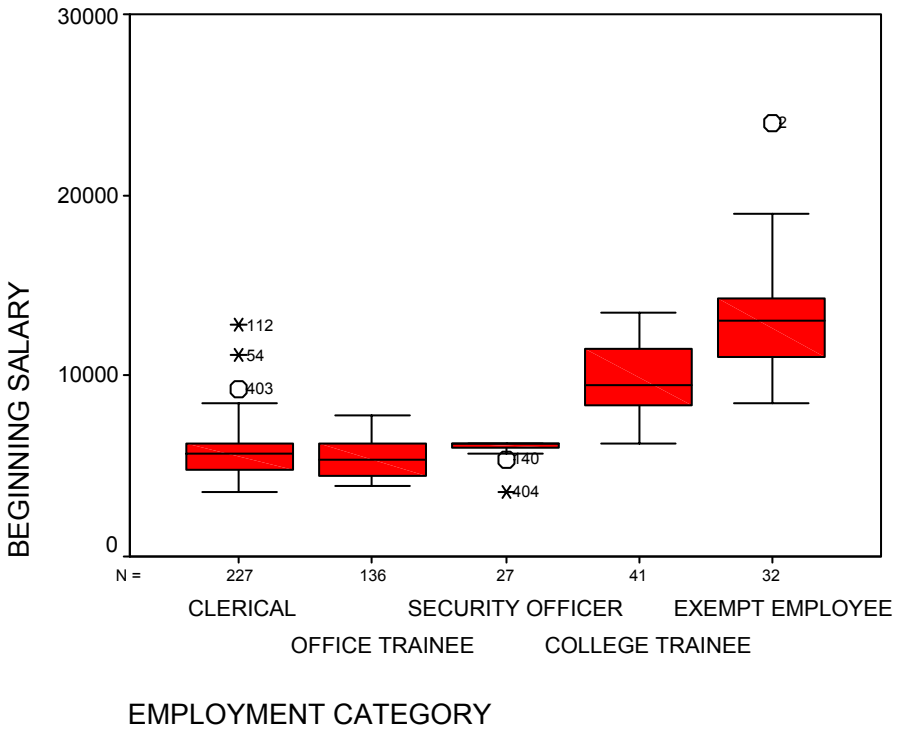
$q(1-\alpha, a, \nu) = q(.95,7,28) = 4.48$

$$t_{crit} = \frac{q_{crit}}{\sqrt{2}} = \frac{4.48}{\sqrt{2}} = 3.17 \text{ (for comparison to } t_{obs})$$

- Sunday $t(28) = 6.09, p < .001$
- Monday $t(28) = 0.31, p = .999$
- Wednesday $t(28) = 3.74, p = .013$
- Thursday $t(28) = 3.06, p = .064$
- Friday $t(28) = 3.23, p = .044$
- Saturday $t(28) = 4.42, p = .002$

- In your manuscript be consistent: report all contrasts as *F*s or report them all as *t*s.

- A return to our Bank starting salary data
  - Before the study was run I decided to compare all groups to the office trainee



  - After looking at the data I want to compare
    - All adjacent pairs
    - (Clerical + Office trainee + Security officer) to
      - College Trainee
      - Exempt Employee
      - (College Trainee + Exempt Employee)

- We previously determined that the variances were not equal in this data, and we will need to take that fact into account.

- Before the study was run (Planned contrasts)
  - I decided to compare all groups to the office trainee
  - We have (a-1) = 4 comparisons, so no correction is necessary
    (Note that many would say to use Bonferroni or Dunnett)

```
ONEWAY salbeg BY jobcat
  /CONTRAST= -1  1  0  0  0
  /CONTRAST= -1 0 1 0 0
  /CONTRAST= -1 0 0 1 0
  /CONTRAST= -1 0 0 0 1.
```

Contrast Tests

		Contrast	Value of Contrast	Std. Error	t	df	Sig. (2-tailed)
BEGINNING SALARY	Assume equal variances	1	-254.98	156.819	-1.626	458	.105
		2	297.16	294.407	1.009	458	.313
		3	4222.54	245.408	17.206	458	.000
		4	7524.93	273.079	27.556	458	.000
	Does not assume equal variances	1	-254.98	116.528	-2.188	345.880	.029
		2	297.16	133.369	2.228	68.848	.029
		3	4222.54	323.084	13.069	46.034	.000
		4	7524.93	562.514	13.377	32.443	.000

- We can simply report the "Does not assume equal variances" results

- Some might argue that we should use a Bonferroni correction

	Dunn/Sidák	Bonferroni
# of tests	$1-(1-.05)^{\frac{1}{c}}$	$.05/c$
4	.0127	.0125

Compare the observed p-value to these adjusted p-values, or compute adjusted p-values

Contrasts 3 and 4 remain significant at the α=.05 level

- After looking at the data I want to compare
  - All adjacent pairs
  - (Clerical + Office trainee + Security officer) to
    - College Trainee
    - Exempt Employee
    - (College Trainee + Exempt Employee)

$\hat{\psi}_1 = (0,-1,1,0,0)$ [We have already tested $(-1,1,0,0,0)$]

$\hat{\psi}_2 = (0,0,-1,1,0)$

$\hat{\psi}_3 = (0,0,0,-1,1)$

$\hat{\psi}_4 = (-1,-1,-1,3,0)$

$\hat{\psi}_5 = (-1,-1,-1,0,3)$

$\hat{\psi}_6 = (-2,-2,-2,3,3)$

- For Contrasts 1-3, we can use Dunnett's T3 (pairwise tests with unequal variances)
- For Contrasts 4-6, we can use the Brown-Forsythe modification of the Scheffe test (complex tests with unequal variances)

```
ONEWAY salbeg BY jobcat
  /CONTRAST= 0  -1  1  0  0
  /CONTRAST= 0  0  -1  1  0
  /CONTRAST= 0 0 0 -1 1
  /CONTRAST= -1 -1 -1 3 0
  /CONTRAST= -1 -1 -1 0 3
  /CONTRAST= -2 -2 -2 3 3.
```

Contrast Tests

		Contrast	Value of Contrast	Std. Error	t	df	Sig. (2-tailed)
BEGINNING SALARY	Assume equal variances	1	552.14	304.697	1.812	458	.071
		2	3925.38	358.432	10.952	458	.000
		3	3302.39	341.131	9.681	458	.000
		4	12625.43	749.105	16.854	458	.000
		5	22532.60	830.832	27.121	458	.000
		6	35158.03	1206.461	29.141	458	.000
	Does not assume equal variances	1	552.14	130.812	4.221	62.580	.000
		2	3925.38	328.506	11.949	48.356	.000
		3	3302.39	637.613	5.179	49.749	.000
		4	12625.43	948.444	13.312	42.235	.000
		5	22532.60	1675.675	13.447	31.542	.000
		6	35158.03	1938.017	18.141	52.405	.000

- Determine the T3 critical value: $q(1\text{-}\alpha, a, \nu)$

  α = Familywise error rate
  a = Number of groups
  ν = the variance adjusted degrees of freedom

$\hat{\psi}_1 \quad t_{obs}(62.58) = 4.22 \quad t_{critical\ adjusted} = \dfrac{q(.05,5,62.58)}{\sqrt{2}} \approx \dfrac{3.98}{\sqrt{2}} = 2.81$

$\hat{\psi}_2 \quad t_{obs}(48.36) = 11.95 \quad t_{critical\ adjusted} = \dfrac{q(.05,5,48.36)}{\sqrt{2}} \approx \dfrac{4.01}{\sqrt{2}} = 2.84$

$\hat{\psi}_3 \quad t_{obs}(49.75) = 5.17 \quad t_{critical\ adjusted} = \dfrac{q(.05,5,49.75)}{\sqrt{2}} \approx \dfrac{4.01}{\sqrt{2}} = 2.84$

- Determine the Brown-Forsythe critical value

$\alpha = .05 \qquad a = 5 \qquad t_{crit} = \sqrt{(a-1)F_{\alpha=.05;a-1,dfw}}$

$\hat{\psi}_4 \quad t_{obs}(42.23) = 13.31 \quad t_{crit} = \sqrt{4F(4,42.23)} = \sqrt{10.37} = 3.22$

$\hat{\psi}_5 \quad t_{obs}(31.54) = 13.45 \quad t_{crit} = \sqrt{4F(4,31.54)} = \sqrt{10.71} = 3.27$

$\hat{\psi}_6 \quad t_{obs}(52.40) = 18.14 \quad t_{crit} = \sqrt{4F(4,52.40)} = \sqrt{10.20} = 3.19$

- All 6 contrasts are significant at $\alpha = .05$.

  When reporting the final results, convert all test statistics to *t*s or *F*s.