Activity #4: ANOVA assumptions, post-hoc tests, linear contrasts

In the last activity, we conducted ANOVA on a few different datasets. Let's take a quick look, once again, at the ANOVA summary tables for a couple of those examples:

Comprehension	SS	df	MS	MSR (F)
Photo	35.05	2	17.53	10.01
Error	94.53	54	1.75	p = 0.0002
Total	129.58	56	MS_{total}	$\eta^2 = 0.27$

Example A: Comprehension of ambiguous prose
Groups = photo before, after, or no photo

Weightloss	SS	df	MS	MSR (F)
Diet	77.6	3	25.87	0.536
Error	4293.7	89	48.244	p = 0.659
Total	4371.3	92	MS_{total}	$\eta^2 = 0.018$

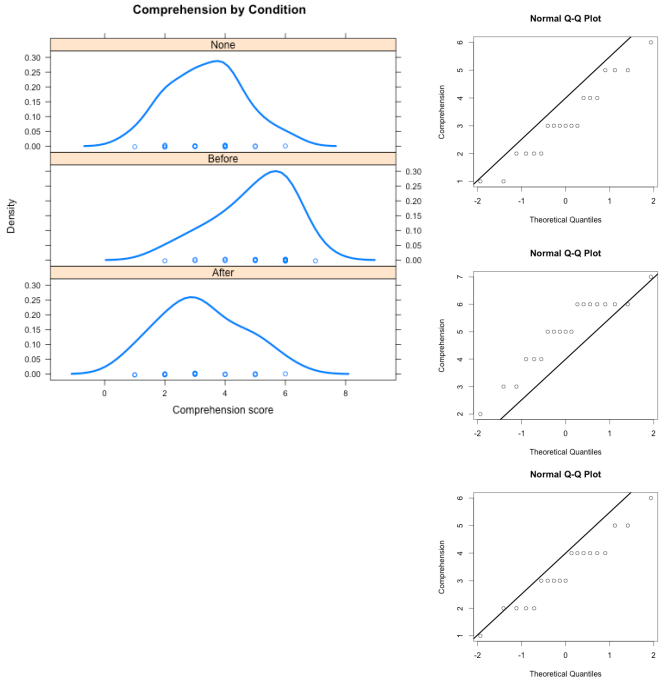
Example B: Pounds lost of various diets
Groups = Atkins, Ornish, WeightWatchers, Zone

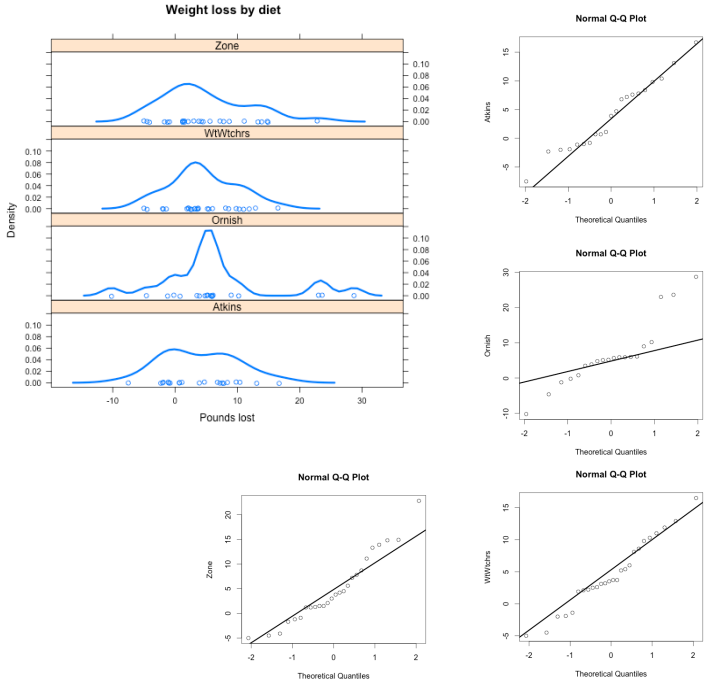
In Example A, we concluded the photo did influence student comprehension of the ambiguous passage.
In Example B, we found no evidence that the diet influenced weight loss.

1. What assumptions did we make in order to conduct the ANOVA? Why did we need to make each assumption?

Assumption	Reason why we need to make this assumption

2. To check for normality, we could look at the distributions of our sample data (histograms, density plots, or Q-Q Plots). Take a look at these plots and determine if you believe the normality assumption has been satisfied.





3. We could also run a test to check for normality. One test, called the Shapiro-Wilk test of normality test, tests the null hypothesis that the data come from a normal distribution. If you want to learn more about this test, you can check out the Wikipedia entry: http://en.wikipedia.org/wiki/Shapiro–Wilk_test

   Below, I've pasted output from the Shapiro-Wilk test for each dataset. What can we conclude?

```
Ambiguous Prose Data
Shapiro-Wilk normality test

data:  ambiguous$Comprehension
W = 0.9433, p-value = 0.009909
```

```
Ambiguous Prose Data
Shapiro-Wilk normality test

data:  diet$WeightLoss
W = 0.9558, p-value = 0.003198
```

4. To test for equal variances, we could conduct an Fmax test by hand. Below, I've conducted an Fmax test for each dataset. What conclusions can you make?

   Ambiguous prose scenario: $F_{\max} = \frac{4.947^2}{3.211^2} = 2.37$ Diet scenario: $F_{\max} = \frac{9.29^2}{5.39^2} = 2.97$

   Fmax table is available at http://bradthiessen.com/html5/stats/m301/4c.pdf

5. We could also run a different test, such as Bartlett's test for equal variances (which, itself, requires a normality assumption). The null hypothesis of this test is that the groups have equal variances. From the output pasted below, what can we conclude?

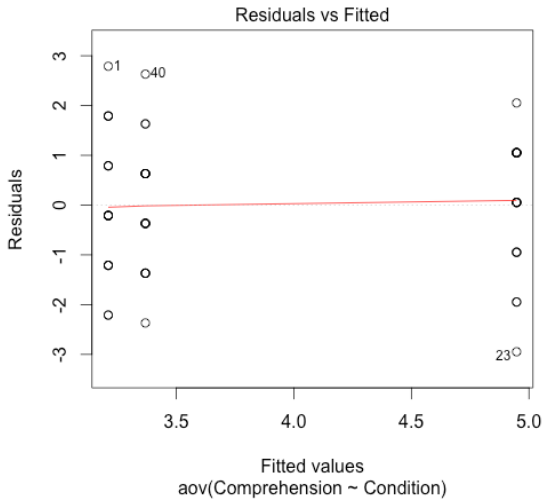
```
Bartlett test of homogeneity of variances

data:  Comprehension by Condition
Bartlett's K-squared = 0.2025, df = 2
p-value = 0.9037
```

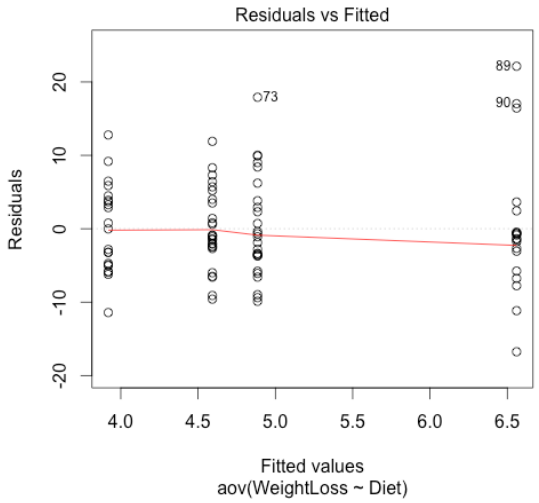
```
Bartlett test of homogeneity of variances

data:  WeightLoss by Diet
Bartlett's K-squared = 7.235, df = 3,
p-value = 0.06477
```

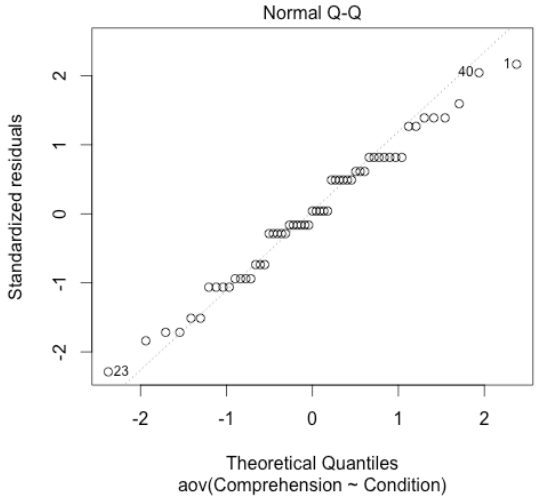
6. Typically, however, we check our model assumptions after we conduct the ANOVA. Write out the models for these two datasets. What did we assume about the parameters of these models?

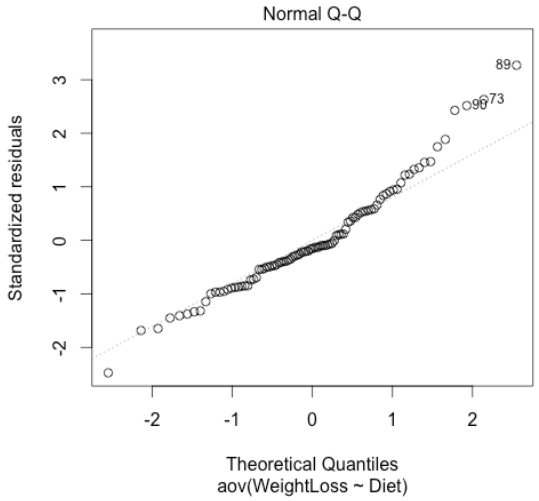
   Comprehension = _____ Weight Loss = _____

7. After conducting an ANOVA on each dataset, I had the computer generate the following plots. Interpret the plots and figure out whether they indicate our assumptions are satisfied:







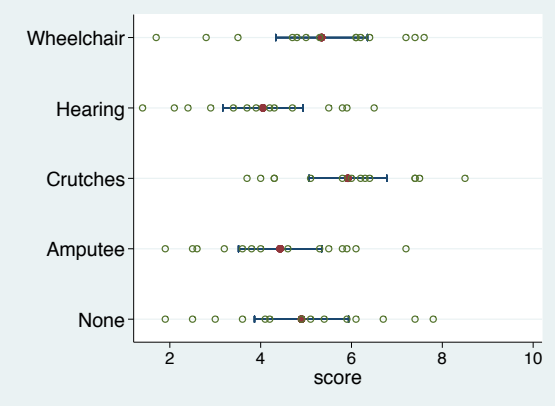


Scenario: In assignment #2, you were introduced to a study designed to determine if disabilities affect perceived performance on job interviews.

In this study, researchers prepared five videotaped job interviews using the same two male actors for each. A set script was designed to reflect an interview with an applicant of average qualifications. The tapes differed only in that the applicant appeared with a different handicap. In one, he appeared in a wheelchair; in a second, he appeared on crutches; in another, his hearing was impaired; in a fourth, he appeared to have one leg amputated; and in the final tape, he appeared to have no handicap

70 undergraduate students from an American university were randomly assigned to view the tapes, 14 to each tape. After viewing the tape, each subject rated the qualifications of the applicant on a 0-10 point scale. Their ratings were as follows:

	No Handicap	Amputee	Crutches	Hearing	Wheelchair
	1.90	1.90	3.70	1.40	1.70
	2.50	2.50	4.00	2.10	2.80
	3.00	2.60	4.30	2.40	3.50
	3.60	3.20	4.30	2.90	4.70
	4.10	3.60	5.10	3.40	4.80
	4.20	3.80	5.80	3.70	5.00
	4.90	4.00	6.00	3.90	5.30
	5.10	4.60	6.20	4.20	6.10
	5.40	5.30	6.30	4.30	6.10
	5.90	5.50	6.40	4.70	6.20
	6.10	5.80	7.40	5.50	6.40
	6.70	5.90	7.40	5.80	7.20
	7.40	6.10	7.50	5.90	7.40
	7.80	7.20	8.50	6.50	7.60
Mean	4.9000	4.4286	5.9124	4.0500	5.3429
StDev	1.7936	1.5857	1.4818	1.5325	1.7483



Source: Cesare, Tannenbaum, Dalessio (1990). *Interviewers' Decisions Related to Applicant Handicap Type and Rater Empathy*. Human Performance, 3(3)

8. From this data, the following ANOVA summary table was produced:

Analysis of Variance

Source	SS	df	MS	F	Prob > F
Between groups	30.5214294	4	7.63035734	2.86	0.0301
Within groups	173.321429	65	2.66648353		
Total	203.842859	69	2.95424433		

Using α=0.05, the researcher rejected the null hypothesis. What conclusions can the researchers make? Could they conclude that a physical disability affects perceived job interview performance?

9. If, after rejecting the null hypothesis in an ANOVA, we want to compare specific treatment means, we'll need to conduct follow-up (*post-hoc*) tests.

   It may make sense to run a series of post hoc tests to compare each possible pair of treatment means. What procedure could we use to test the difference between two group means:

10) In this example (with 5 treatments), how many pairwise comparisons could we possibly make? How many pairwise comparisons could we make if we had *g* groups?

11) As we've already discussed, running multiple tests on the same set of data increases our overall α-error rate. If we set α=0.05 and conducted all 10 t-tests, what would our overall (familywise) α-error rate be? What would be our familywise α-error rate if we used α=*a* to conduct all possible pairwise comparisons from *g* groups?

12) Suppose we really want our familywise α-error rate to be 0.05. At what level must we set α for each of our 10 tests?

This is the idea behind the *Bonferroni adjustment*. In order to control the familywise α-error rate, we need to divide that α-level by the number of tests we intend to conduct.

Bonferroni adjustment: $\alpha_{\text{each test}} = \dfrac{\alpha_{\text{familywise}}}{\#\text{ of tests}}$

13) We then use that adjusted α-value to conduct independent samples t-tests. As you recall, the formula for a t-statistic is:

$$t_{df} = \frac{\text{observed - hypothesized}}{\text{standard error}} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{s^2_{\text{pooled}}}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

We'll use this formula with a slight modification. Rather than calculate the pooled standard deviation, we can use something from our ANOVA summary table. In this study, what is our best estimate for the pooled standard deviation of our treatments?

Bonferroni method to compare two treatment means following an ANOVA:

14) Let's use the Bonferroni method to compare all possible pairs of means in our study. To do this, we need the following:

	No Handicap	Amputee	Crutches	Hearing	Wheelchair
n	14	14	14	14	14
Mean	4.9000	4.4286	5.9124	4.0500	5.3429
StDev	1.7936	1.5857	1.4818	1.5325	1.7483

MSE	2.6665

First, let's compare the **no handicap** treatment to the **amputee** treatment. Calculate the t-statistic, determine the appropriate α-level, and make a decision.

t = _____ =

α-level = _____

decision/conclusion = _____

Note: To find the critical t-value (or to estimate the p-value), use the online calculators linked from the class website.

15) Let's do this one more time, let's compare **crutches** to **hearing**. Do you see why this might be the first pairwise comparison we would want to make? Run the test and state your conclusion.

16) This method could quickly become tedious. Thankfully, computers can run these tests very quickly. Also, we could choose to calculate confidence intervals (instead of t-tests) and display all pairwise comparisons in a single table:

Comparison	Difference	Standard Error	t-critical	Confidence Interval	Significant?
	$\bar{X}_1 - \bar{X}_2$	$\sqrt{MS_E\left(1/n_1 + 1/n_2\right)}$	$t_{df_E}^{\alpha=\alpha/2(\text{tests})}$	Diff ± t(SE)	How do we determine?
None vs. Amputee	0.4174	0.61714	2.91	(-1.378 , 2.267)	_____
None vs. Crutches	-1.0214	0.61714	2.91	(-2.82 , 0.774)	_____
None vs. Hearing	0.85	0.61714	2.91	(-0.95 , 2.65)	_____
None vs. Wheelchair	-0.4429	0. 61714	2.91	(-2.24 , 1.35)	_____

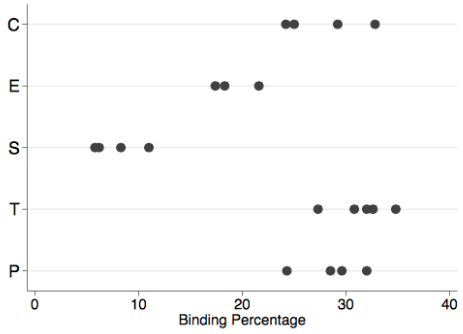
17) Here's the output from a computer program called *Stata*. What conclusions can you make?

Comparison of score by handicap
(Bonferroni)

Row Mean- Col Mean	None	Amputee	Crutches	Hearing
Amputee	-.471429 1.000			
Crutches	1.02143 1.000	1.49286 0.184		
Hearing	-.85 1.000	-.378571 1.000	-1.87143 0.035	
Wheelcha	.442857 1.000	.914286 1.000	-.578572 1.000	1.29286 0.401

18) The following table displays the binding percentages of 5 different types of drugs. Lower numbers are better:

	Measurements					Mean	Std. Dev.	Sample
(P) Penicillin G	29.6	24.3	28.5	32.0		28.600	3.2177	$n_1 = 4$
(T) Tetracycline	27.3	32.6	30.8	34.8	32.0	31.500	2.7604	$n_2 = 5$
(S) Streptomycin	05.8	06.2	11.0	08.3		7.825	2.3838	$n_3 = 4$
(E) Erythromycin	21.6	17.4	18.3			19.100	2.2113	$n_4 = 3$
(C) Chloramphenicol	29.2	32.8	25.0	24.2		27.800	3.9900	$n_5 = 4$
				Total group:		M = 23.585	$s_t = 9.3889$	N = 20



Number of obs = 20 R-squared = 0.9187
Root MSE = 3.0125 Adj R-squared = 0.8971

Source	Partial SS	df	MS	F	Prob > F
drug	1538.75794	4	384.689486	42.39	0.0000
Residual	136.1275	15	9.07516666		
Total	1674.88544	19	88.1518654		

What conclusions can you make from the computer output (from *SPSS)* printed on the top of the next page?

Multiple Comparisons

Dependent Variable: BINDING

Bonferroni

(I) DRUG	(J) DRUG	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
P	T	-2.900	2.0208	1.000	-9.541	3.741
	S	20.775*	2.1302	.000	13.775	27.775
	E	9.500*	2.3008	.009	1.939	17.061
	C	.800	2.1302	1.000	-6.200	7.800
T	P	2.900	2.0208	1.000	-3.741	9.541
	S	23.675*	2.0208	.000	17.034	30.316
	E	12.400*	2.2000	.000	5.171	19.629
	C	3.700	2.0208	.870	-2.941	10.341
S	P	-20.775*	2.1302	.000	-27.775	-13.775
	T	-23.675*	2.0208	.000	-30.316	-17.034
	E	-11.275*	2.3008	.002	-18.836	-3.714
	C	-19.975*	2.1302	.000	-26.975	-12.975
E	P	-9.500*	2.3008	.009	-17.061	-1.939
	T	-12.400*	2.2000	.000	-19.629	-5.171
	S	11.275*	2.3008	.002	3.714	18.836
	C	-8.700*	2.3008	.018	-16.261	-1.139
C	P	-.800	2.1302	1.000	-7.800	6.200
	T	-3.700	2.0208	.870	-10.341	2.941
	S	19.975*	2.1302	.000	12.975	26.975
	E	8.700*	2.3008	.018	1.139	16.261

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

19) If we conduct an ANOVA and reject the null hypothesis, will we always find at least one significant pairwise difference?

20) If we conduct an ANOVA and retain the null hypothesis, is it possible for us to find at least one significant pairwise difference?

21) Let's turn our attention back to the study about disabilities:

	No Handicap	Amputee	Crutches	Hearing	Wheelchair
n	14	14	14	14	14
Mean	4.9000	4.4286	5.9124	4.0500	5.3429
StDev	1.7936	1.5857	1.4818	1.5325	1.7483

MSE	2.6665

Suppose we're interested in a seemingly simple question: Does having a disability affect perceived job interview performance? To address this question, we could run 4 t-tests using a Bonferroni adjustment (comparing the *no handicap* group to each of the disability groups).

Another way to address this question would be to calculate a linear combination of our treatment means using the *Scheffé method*.

In this method, we first need to calculate a *contrast* (a linear combination of means):

$$\psi = c_1\mu_1 + c_2\mu_2 + c_3\mu_3 + \ldots + c_2\mu_2 \qquad \text{where } \sum c = 0$$

Choose constants in order to compare the **no handicap** group to all four disability groups. Calculate this contrast:

$$\hat{\psi} =$$

Under a null hypothesis, what would we expect the value of this contrast to be?

$$\text{If } H_0 \text{ is true, then } E[\psi] =$$

Our contrast isn't zero, but we wouldn't expect it to be. Remember, an expected value tells us the average value we would expect in the long-run. Even if our treatment means did not differ (in the population), we'd expect our sample means to differ and our sample contrast to be nonzero.

So our task is to determine the likelihood of observing a contrast as or more extreme than the one we calculated (assuming the null hypothesis is true). To do this, we need to figure out the sampling distribution of our contrast.

Recall our general formula for a t-test. We may be able to use this formula to test the significance of our contrast.

$$t = \frac{\text{observed - hypothesized}}{\text{standard error}} = \frac{\hat{\psi} - 0}{SE_\psi} = \frac{\hat{\psi}}{SD_\psi / \sqrt{n}} = \frac{\hat{\psi}}{\sqrt{\text{var}_\psi} / \sqrt{n}}$$

If this works, we only need to learn how to calculate the standard error of a contrast. Then, we'd have our test statistic and we could conduct the test and estimate a p-value.

To derive the formula for the standard error, let's begin by looking at our contrast in this example:

$$\hat{\psi} = (4)\bar{X}_1 + (-1)\bar{X}_2 + (-1)\bar{X}_3 + (-1)\bar{X}_4 + (-1)\bar{X}_5$$

Note that this is one of an infinite number of contrasts we could have used to address our research question. If we assume our groups are independent (which we assumed when we conducted the ANOVA), then we could calculate:

$$\text{var}(\psi) = \text{var}(4\bar{X}_1) + \text{var}(-1\bar{X}_2) + \text{var}(-1\bar{X}_3) + \text{var}(-1\bar{X}_4) + \text{var}(-1\bar{X}_5)$$

You may recall from a previous statistics class that $\text{var}(ax) = a^2 \text{ var}(x)$. Using this, we can simplify:

$$\text{var}(\psi) =$$
$$= \text{var}\left(c_1\bar{X}_1\right) + \text{var}\left(c_2\bar{X}_2\right) + \text{var}\left(c_3\bar{X}_3\right) + \text{var}\left(c_4\bar{X}_4\right) + \text{var}\left(c_5\bar{X}_5\right)$$
$$= c_1^2 \text{ var}\left(\bar{X}_1\right) + c_2^2 \text{ var}\left(\bar{X}_2\right) + c_3^2 \text{ var}\left(\bar{X}_3\right) + c_4^2 \text{ var}\left(\bar{X}_4\right) + c_5^2 \text{ var}\left(\bar{X}_5\right)$$

You may also recall that from the Central Limit Theorem: $\text{var}(\bar{X}) = \frac{\sigma_x^2}{n}$

$$= c_1^2 \frac{\sigma_1^2}{n_1} + c_2^2 \frac{\sigma_2^2}{n_2} + c_3^2 \frac{\sigma_3^2}{n_3} + c_4^2 \frac{\sigma_4^2}{n_4} + c_5^2 \frac{\sigma_5^2}{n_5}$$
$$= \left(c_1^2 + c_2^2 + c_3^2 + c_4^2 + c_5^2\right)\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} + \frac{\sigma_3^2}{n_3} + \frac{\sigma_4^2}{n_4} + \frac{\sigma_5^2}{n_5}\right)$$

If we're assuming the variances within each group are equal, we can simply this formula to:

$$= \sum \frac{c_j^2}{n_j} \sigma_x^2$$

Finally, our best estimate of the pooled (weighted average) variance is MSE, so we can write:

$$\text{var}(\psi) = MS_E \sum \frac{c_j^2}{n_j}$$

We can now substitute this value into our test statistic:

$$\frac{\text{observed - hypothesized}}{\text{standard error}} = \frac{\hat{\psi}}{SE_\psi} = \frac{\hat{\psi}}{\sqrt{MS_E \sum \frac{c_j^2}{n_j}}} \sim \sqrt{a-1(F)}$$

22) Use the *Scheffé method* to compare the **no handicap** group to the other disability groups. What conclusions can you make?

23) Use the *Scheffé method* to compare the **no handicap and crutches** group to the other disability groups. What conclusions can you make?

Scenario: To what degree do different teaching methods affect student achievement? To address this question, four different teaching methods were used in 5th grade classrooms in and around St. Louis. Three of these methods were specially designed to deal with student heterogeneity within the typical classroom. The fourth method was a traditional textbook-based method with very little provision for individual differences except for the range of difficulty of the exercises.

A1) Missouri Mathematics Program (MMP): A high ratio of active teaching to seatwork, frequent feedback, smooth transitions between topics, and other features derived from an analysis of the practices of outstanding teachers.

A2) Ability-Grouped Active Teaching (AGAT): Incorporates major features of the Missouri program, but modified to accommodate to classrooms in which two ability groups (top 60%; lower 40%) have been organized.

A3) Team Assisted Individualization (TAI): Classes organized into heterogeneous 5-member learning groups. Individualized lesson materials, largely self-pacing, are assigned to each member of the basis of his/her ability. Students on each team help their teammates and take responsibility for checking work. At the end of each week, teams that meet pre-set criteria receive attractive certificates and other forms of recognition.

A4) Textbook-Based Method (TBM): Traditional, teacher-centered method with undifferentiated assignments and no specific activities or methods designed to meet the specific learning rates of individual students.

Pupils were assigned randomly to methods within the constraints imposed by the sizes of the school districts, attendance center boundaries, teacher assignments, etc.

To measure student achievement, students were administered the *Math Concepts & Estimation* subtest of the Comprehensive Tests of Basic Skills (scores represent grade equivalent units).

Here is the data and an ANOVA summary table from the study:

Method	Subjects	Mean Score	Std. Deviation
MMP	162	5.70	1.16
AGAT	98	6.43	1.29
TAI	114	6.21	1.27
TBM	106	5.65	1.07
Total	480	5.96	

Source	Sums of Squares	df	Mean Square	Mean Square Ratio
SSA	49.911	3	16.64	11.64
SSE	680.5315	476	1.43	Fcv = 2.60
Total	730.4425	479	Reject null hypothesis. There are group differences (but we don't know where)	

24) Suppose we were going to test all possible pairs of group means. Use the Bonferroni adjustment to test if the traditional method versus the TAI method. What conclusions can you make?

25) Use the *Scheffé* method to compare the traditional method versus the other methods. What conclusions can you make?

Scenario: 45 female dog lovers were asked to do a stressful task. Randomly, the subjects were assigned to one of the following groups:

Control Group: This group of subjects completed the task alone
Friend Group: This group of subjects completed the task in the company of a friend
Dog Group: This group of subjects completed the task in the company of their dog

Stress levels while completing the task were rated on a scale from 0-100. Here are the results of this study:

Group	Subjects	Mean Score	Std. Deviation
Control	15	82.524	9.242
Friend	15	91.325	8.341
Dog	15	73.483	9.97

Source	Sums of Squares	df	Mean Square	Mean Square Ratio
SSA	2387.7	2	1193.8	14.08
SSE	3561.3	42	84.8	$p < 0.01$
Total	5949	44	Reject null hypothesis. There are group differences (but we don't know where)	

26) Use the *Bonferroni* method to compare the control group against the dog group. What conclusions can you make?

27) Use the *Scheffé* method to compare the control group versus the other two groups. What conclusions can you make?

See the R Code for this activity to see how to conduct these post-hoc tests on a computer