Multiple/Post Hoc Group Comparisons in ANOVA

Note: We may just go over this quickly in class. The key thing to understand is that, when trying to identify where differences are between groups, there are different ways of adjusting the probability estimates to reflect the fact that multiple comparisons are being made.

Introduction. In a one-way ANOVA, the F statistic tests whether the treatment effects are all equal, i.e. that there are no differences among the means of the J groups. A significant F value indicates that there are differences in the means, but it does not tell you where those differences are, e.g. group 1's mean might be different than group 2's mean but not different from group 3's mean.

To isolate where the differences are, you could do a series of pairwise T-tests. The problem with this is that the significance levels can be misleading. For example, if you have 7 groups, there will be 21 pairwise comparisons of means; if using the .05 level of significance, you would expect at least one statistically significant difference even if no differences exist.

Therefore, various methods have been developed for doing multiple comparisons of group means. In SPSS, one way to accomplish this is via the use of the /POSTHOC parameter on the Oneway command. We'll present the SPSS output and then explain what the different parts mean.

```
ONEWAY
  score BY program
  /STATISTICS DESCRIPTIVES
  /MISSING ANALYSIS
  /POSTHOC =  LSD BONFERRONI SIDAK SCHEFFE ALPHA(.05).
```

Oneway

Descriptives

SCORE

					95% Confidence Interval for Mean			
	N	Mean	Std. Deviation	Std. Error	Lower Bound	Upper Bound	Minimum	Maximum
1.00	5	11.8000	1.92354	.86023	9.4116	14.1884	9.00	14.00
2.00	5	8.8000	1.64317	.73485	6.7597	10.8403	6.00	10.00
3.00	5	12.2000	1.30384	.58310	10.5811	13.8189	11.00	14.00
4.00	5	8.6000	1.51658	.67823	6.7169	10.4831	7.00	11.00
Total	20	10.3500	2.25424	.50406	9.2950	11.4050	6.00	14.00

ANOVA

SCORE

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	54.950	3	18.317	7.045	.003
Within Groups	41.600	16	2.600		
Total	96.550	19			

Post Hoc Tests

Multiple Comparisons

Dependent Variable: SCORE

	(I) PROGRAM	(J) PROGRAM	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
LSD	1.00	2.00	3.0000*	1.01980	.009574	.8381	5.1619
		3.00	-.4000	1.01980	.700062	-2.5619	1.7619
		4.00	3.2000*	1.01980	.006355	1.0381	5.3619
	2.00	1.00	-3.0000*	1.01980	.009574	-5.1619	-.8381
		3.00	-3.4000*	1.01980	.004207	-5.5619	-1.2381
		4.00	.2000	1.01980	.846988	-1.9619	2.3619
	3.00	1.00	.4000	1.01980	.700062	-1.7619	2.5619
		2.00	3.4000*	1.01980	.004207	1.2381	5.5619
		4.00	3.6000*	1.01980	.002781	1.4381	5.7619
	4.00	1.00	-3.2000*	1.01980	.006355	-5.3619	-1.0381
		2.00	-.2000	1.01980	.846988	-2.3619	1.9619
		3.00	-3.6000*	1.01980	.002781	-5.7619	-1.4381
Bonferroni	1.00	2.00	3.0000	1.01980	.057442	-.0679	6.0679
		3.00	-.4000	1.01980	1.000000	-3.4679	2.6679
		4.00	3.2000*	1.01980	.038130	.1321	6.2679
	2.00	1.00	-3.0000	1.01980	.057442	-6.0679	.0679
		3.00	-3.4000*	1.01980	.025242	-6.4679	-.3321
		4.00	.2000	1.01980	1.000000	-2.8679	3.2679
	3.00	1.00	.4000	1.01980	1.000000	-2.6679	3.4679
		2.00	3.4000*	1.01980	.025242	.3321	6.4679
		4.00	3.6000*	1.01980	.016686	.5321	6.6679
	4.00	1.00	-3.2000*	1.01980	.038130	-6.2679	-.1321
		2.00	-.2000	1.01980	1.000000	-3.2679	2.8679
		3.00	-3.6000*	1.01980	.016686	-6.6679	-.5321
Sidak	1.00	2.00	3.0000	1.01980	.056084	-.0575	6.0575
		3.00	-.4000	1.01980	.999272	-3.4575	2.6575
		4.00	3.2000*	1.01980	.037530	.1425	6.2575
	2.00	1.00	-3.0000	1.01980	.056084	-6.0575	.0575
		3.00	-3.4000*	1.01980	.024978	-6.4575	-.3425
		4.00	.2000	1.01980	.999987	-2.8575	3.2575
	3.00	1.00	.4000	1.01980	.999272	-2.6575	3.4575
		2.00	3.4000*	1.01980	.024978	.3425	6.4575
		4.00	3.6000*	1.01980	.016571	.5425	6.6575
	4.00	1.00	-3.2000*	1.01980	.037530	-6.2575	-.1425
		2.00	-.2000	1.01980	.999987	-3.2575	2.8575
		3.00	-3.6000*	1.01980	.016571	-6.6575	-.5425
Scheffe	1.00	2.00	3.0000	1.01980	.068155	-.1789	6.1789
		3.00	-.4000	1.01980	.984100	-3.5789	2.7789
		4.00	3.2000*	1.01980	.048181	.0211	6.3789
	2.00	1.00	-3.0000	1.01980	.068155	-6.1789	.1789
		3.00	-3.4000*	1.01980	.033774	-6.5789	-.2211
		4.00	.2000	1.01980	.997930	-2.9789	3.3789
	3.00	1.00	.4000	1.01980	.984100	-2.7789	3.5789
		2.00	3.4000*	1.01980	.033774	.2211	6.5789
		4.00	3.6000*	1.01980	.023519	.4211	6.7789
	4.00	1.00	-3.2000*	1.01980	.048181	-6.3789	-.0211
		2.00	-.2000	1.01980	.997930	-3.3789	2.9789
		3.00	-3.6000*	1.01980	.023519	-6.7789	-.4211

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

We have seen the descriptive statistics and the ANOVA table before, so we will focus on the Posthoc comparisons table.

Mean difference. This column gives the difference in the means of the 2 groups. For example, group 1's mean is 11.8, group 2's mean is 8.8, so the difference is 3. An asterisk by the value indicates whether the difference is statistically significant given the method of multiple comparisons being used. (More on methods below.)

Standard error. In a One-way Anova, the standard error of the difference between the two means of groups i and j is

$$s_{\hat{\mu}_i - \hat{\mu}_j} = \sqrt{MSE * \left(\frac{1}{N_i} + \frac{1}{N_j} \right)}$$

(Recall that MSE is another name for MS Within.) In this particular example, the group sizes are all the same, which is why the reported standard errors are all the same, but this will not be true when group sizes differ. In this example,

$$s_{\hat{\mu}_i - \hat{\mu}_j} = \sqrt{MSE * \left(\frac{1}{N_i} + \frac{1}{N_j} \right)} = \sqrt{2.6 * \left(\frac{1}{5} + \frac{1}{5} \right)} = \sqrt{1.04} = 1.0198$$

Sig. This column gives you the significance of the difference under the multiple comparison method being used. To understand this, we need explain each of the 4 methods being used and what their rationale is.

LSD. LSD stands for Least Significant Difference t test. This test does *not* control the overall probability of rejecting the hypotheses that some pairs of means are different, while in fact they are equal, i.e. it doesn't matter if you are comparing 1 pair of means or a 100, no adjustment is made for the number of comparisons. The formula is

$$LSD_{i-j} = \frac{\hat{\mu}_i - \hat{\mu}_j}{s_{\hat{\mu}_i - \hat{\mu}_j}}$$

This statistic has a T distribution with N-J d.f. where J = number of groups. So, for example, the LSD value for the comparison of groups 1 and 2 is

$$LSD_{i-j} = \frac{\hat{\mu}_i - \hat{\mu}_j}{s_{\hat{\mu}_i - \hat{\mu}_j}} = \frac{3}{1.0198} = 2.94175, \text{ d.f} = 16$$

The 2-tailed probability of getting a t value this large or larger in magnitude if the null is true is only .009574, i.e. there is less than 1 chance in a hundred that their could be no difference in group means and the sample would produce a difference in means that is this large.

Alternatively, you can just square the LSD statistic; the resulting value has an F distribution with d.f. 1, N-J, i.e.

$$LSD^2{}_{i-j} = \left(\frac{\hat{\mu}_i - \hat{\mu}_j}{s_{\hat{\mu}_i - \hat{\mu}_j}} \right)^2 = \left(\frac{3}{1.0198} \right)^2 = 8.654, \text{ d.f} = 1, 16$$

The name LSD derives from the fact that you determine what the smallest difference between means is that would be statistically significant. If the actual difference is greater than that, then you regard the result as statistically significant. In this case, note that if we are doing a two-tailed test using the .01 level of significance, the critical value for a t with 16 d.f. is 2.921. (For an F with d.f. 1, 16, the critical value is 8.53) For the.05 level (which is what we told ONEWAY to use) the critical value is 2.12, hence there is an * by the value of 3 in the mean difference column.

Note that LSD makes no adjustment for the fact that multiple comparisons are being made. In this case, there are 6 possible pairwise comparisons; hence the odds that at least one of them would be significant at the .05 level (even if there are no differences) is actually much greater than .05, i.e. if you do enough comparisons, just by chance some will show up as significant. The remaining methods offer different ways of adjusting the significance levels to compensate for this.

Bonferroni. The Bonferroni adjustment is the simplest. It basically multiplies each of the significance levels from the LSD test by the number of tests performed, i.e. J*(J-1)/2. If this value is greater than 1, then a significance level of 1 is used. So, for example, the LSD test reports that the difference between groups 1 and 2 is significant at the .009574 level. The Bonferroni adjustment multiplies this by 6 (the number of pairwise comparisons when there are 4 groups) and reports a significance level of 6 * .009574 = .057442. Note that this is greater than .05, so the difference between groups 1 and 2 is not considered significant (hence no * in the mean difference column).

For group 1 versus 3, LSD reports that the difference is only significant at the .7 level. Since 6 * .7 is greater than 1, the Bonferroni adjustment reports a significance level of 1. If you compare the significance levels of LSD and Bonferroni, you'll see that Bonferroni is always 6 times larger than LSD, or else 1, i.e. Bonferroni = Minimum(6*LSD, 1).

An additional implication of this is that LSD results have to be significant at the .05/6 = .00833 level in order to be significant at the .05 level under Bonferroni. Similarly, if we had 7 groups and hence 21 pairwise comparisons, the LSD test would have to be significant at the .05/21 = .00238 level to be significant after the Bonferroni adjustment.

Sidak. While simple, the Bonferroni adjustment actually overcompensates for the fact that multiple comparisons are being made, e.g. if you do 21 tests, the probability is NOT 1.05 that at least one of them will be significant at the .05 level; rather, it is $1 - .95^{21} = .659$. The Sidak adjustment computes the level of significance as

$$1-(1-LSD_{significance})^{J*(J-1)/2}$$

So, for example, for the group 1 versus group 2 comparison, the Sidak significance is

$$1-(1-LSD_{significance})^{J*(J-1)/2} = 1-(1-.009574)^{6} = .056087$$

This is a little more significant than what Bonferroni came up with but still more than .05, so the difference between groups 1 and 2 is not considered significant. For group 1 versus 3,

$$1-(1-LSD_{significance})^{J*(J-1)/2} = 1-(1-.700062)^{6} = .999272$$

Scheffe. The Scheffe test takes a somewhat different approach. The Scheffe test computes an F statistic with d.f. = J-1, N-J.

Scheffe = $LSD^2/(J - 1)$.

So, for group 1 versus group 2, the Scheffe value is 8.654/3 = 2.8847. An F value of 2.8847 with d.f. = 3, 16 is significant at the .0682 level. (For an F with d.f. 3, 16, the test statistic has to be 3.01 or larger to be significant at the .05 level). Again, Scheffe says the group 1 versus group 2 difference is not significant at the .05 level.

Confidence Intervals. I won’t go into the details of how the confidence intervals are computed. But, note that, if 0 falls within the confidence interval, you should NOT reject the null hypothesis that there is no difference in the means.

Other Comments.

- I’m not sure that it makes a whole lot of difference which of the adjustment methods you use. But, since all 3 of these will show up in the literature, you should understand the general idea that these are methods designed to reduce our overall chance of falsely rejecting each hypothesis to α rather than letting it increase with each additional test.
- The flip side is that the adjustment methods increase the likelihood we will stick with the null when we should reject it. For example, suppose there were 7 groups and each pairwise difference was significant at the .04 level. It is extremely unlikely that you would get so many significant differences by chance and your overall F value in your ANOVA would be highly significant. Nonetheless, if you made the Bonferroni adjustment, each would now be significant at the .84 level and hence none of the differences would be considered significant.
- Similar adjustments can be done in other contexts, e.g. in a correlation matrix, some correlations can be significant just by chance; so, you’ll sometimes see Bonferroni or other adjustments being made.
- I’ve never been asked to make any such adjustment in my work! Indeed, it gets complicated to do so once you get beyond a one-way Anova framework. But, such adjustments are probably much more common in other fields of study.