Post-hoc Testing

After running an ANOVA researchers often wish to go back determine which pairs of conditions are actually different. Recall, the null hypothesis for a one-way ANOVA is that the (population) means are equal. For instance, if we have three groups the null and alternative hypotheses are . . .

$$H_0 : \mu_1 = \mu_2 = \mu_3 \qquad H_A : \text{the means are different}$$

A statistically significant ANOVA results, doesn't tell us which conditions are different, only that they are not all equal. When these tests are done after we've looked at the data these are called *post-hoc* tests. For reasons that we will discuss in a moment, it is a bad idea to just run t-tests on all pairs of conditions. Instead there are two common ways of doing post-hoc tests: Tukey's Honestly Significant Difference and the Scheffe Test.

Tukey's Honestly Significant Difference

The key idea with Tukey's honestly significant difference is to determine a minimum threshold for differences between pairs. If the (absolute value) of the difference between pairs is larger than this threshold then we say that the difference is significant (honestly). To calculate the HSD we use the $MS = SS/df$ from the denominator of the F ratio and a new statistic q the "studentized range statistic," which is determined by the α, degrees of freedom, and number of conditions.

For independent measures: $$HSD = q\sqrt{\frac{MS_{\text{within groups}}}{n}}$$

For dependent measures: $$HSD = q\sqrt{\frac{MS_{\text{error}}}{n}}$$

We then compare all possible pairs of differences between the group means to this minimum, and any differences greater than the HSD are considered to be different pairs.

Scheffe Test

The Scheffe Test, rather than comparing the differences, instead treats each pair-wise comparison as a separate hypothesis test. Remember, the problem of comparing two group means is just a special case of one-way ANOVA. So we can use an F-test, but we'll make a small adjustment to focus on the two groups in question.

For independent-measures: $$F = \frac{SS^*_{\text{between}}/df_{\text{between}}}{SS_{\text{within}}/df_{\text{within}}}$$

For dependent-measures: $$F = \frac{SS^*_{\text{between}}/df_{\text{between}}}{SS_{\text{error}}/df_{\text{error}}}$$

All of the values in this calculation are straight from the full ANOVA except for SS_{between^*}. Instead of using the original formula we focus on two groups of interest i and j:

$$SS^*_{\text{between}} = n_i(\bar{X}_i - \bar{X}_{\text{All}})^2 + n_j(\bar{X}_j - \bar{X}_{\text{All}})^2$$

Then if F exceeds the critical value F_c we conclude that the two groups i and j are different.

Correcting for Multiple Comparisons

The reason that these adjustments are necessary (rather than using just a standard t-test) is that we can't just go around testing everything for differences. The error rate, depends on how many comparisons we make. Recall that α determines the probability of a Type I error - with $\alpha = 0.05$ we will have a false positive 1 time out of 20. If we were to do 20 t-tests we expect one of them to result in a statistically significant difference even if there is absolutely no effect. In this situation the *test-wise* errors add up so that the *family-wise* error rate (the probability that at least one comparison is significant) is actually much larger than α. For example, in a one-way ANOVA with $k = 3$ there are 3 possible comparisons: 1-2, 1-3, and 2-3. In the worst case, the probability that at least one comparison results in a statistically significant difference is 3α instead of just α. In general, there are $k(k-1)/2$ possible comparisons for a one-way ANOVA, and the family-wise error rate gets larger as we make more comparisons.

Tukey's HSD and the Scheffe test are both designed to be slightly more conservative than a post-hoc t-test, and attempt to correct for the problem of *multiple comparisons*. While HSD and the Scheffe test are popular for behavioral experiments with relatively few experimental conditions, in other areas of psychology, such as cognitive neuroscience, we might be making many more comparisons. In an fMRI experiment we might want to see which voxels (3D pixels in the data) are correlated with a specific stimulus while a subject is in the scanner. Since we may have over 1 million voxels!, we need to correct for the fact that many of these correlations (~50,000) will be statistically significant just by chance. In such large-scale experiments, there are two common ways of correcting for multiple comparisions:

1. Bonferonni Correction. Instead of evaluating the tests with α, we can lower the test-wise error rate proportionally to the number of comparisons. After doing a one-way ANOVA with $k = 3, \alpha = 0.05$ we can test pairs of conditions using t-tests with $\alpha^* = 0.05/3$. In general, if we make m comparisions and want a family-wise error rate $\leq \alpha$ we can use a test-wise error rate of α/m.

2. False Discovery Rate (FDR). In many cases, Bonferonni correction is overly conservative and can result in more false negatives than desired. Recently, several methods have been developed for more liberally adusting the FDR. The most common of these procedures (Benjamini-Hochberg) is to collect the p-values for all m tests, sort them from small to large, and then only reject the first k null hypotheses that satisfy $P_{(k)} \leq \frac{k}{m}\alpha$.