Multiple Comparisons

- Most significance tests are designed to be carried out in isolation
- But if you do a lot of tests and all the null hypotheses are true, the chance of rejecting at least one of them can be a lot more than 0.05. This is **inflation of the Type I error rate**.
- Multiple comparisons (follow-up tests, post hoc tests, probing) offer a solution.

Multiple comparisons

- Protect a *family* of tests against Type I error at some *joint* significance level (usually 0.05)
- If all the null hypotheses are true, the probability of rejecting at least one is no more than 0.05

Multiple comparisons of contrasts in a one-way design: Assume all means are equal in the population

- Bonferroni
- Tukey
- Scheffé

Bonferroni

- Applies to *any* collection of k tests
- Do the tests as usual
- Reject each H_0 if $p < 0.05/k$
- Or, adjust the p-values. Multiply them by k, and reject if the adjusted $p < 0.05$

Bonferroni

- Advantage: Flexibility
- Advantage: Easy to do

- Disadvantage: Must know what all the tests are before seeing the data
- Disadvantage: A little conservative; the true joint significance level is *less* than 0.05.

Tukey (HSD)

- Applies only to pairwise comparisons of means
- If sample sizes are equal, it's most powerful, period
- If sample sizes are not equal, it's a bit conservative

Scheffé

- Find the usual critical value for the initial test. Multiply by p-1. This is the Scheffé critical value.
- Family includes *all* contrasts
- You don’t need to specify them in advance

Scheffé

- Follow-up tests cannot be significant if the initial overall test is not. Not quite true of Bonferroni and Tukey.
- If the initial test (of p-1 contrasts) is significant, there is a single contrast that is significant (not necessarily a pairwise comparison)
- Adjusted p-value is the tail area beyond F times (p-1)

Which method should you use?

- If the sample sizes are nearly equal and you are only interested in pairwise comparisons, use Tukey because it's most powerful

- If the sample sizes are not close to equal and you are only interested in pairwise comparisons, there is (amazingly) no harm in applying all three methods and picking the one that gives you the greatest number of significant results.

- If you are interested in contrasts that go beyond pairwise comparisons and you can specify *all* of them before seeing the data, Bonferroni is almost always more powerful than Scheffé. (Tukey is out.)

- If you want lots of special contrasts but you don't know exactly what they all are, Scheffé is the only honest way to go, unless you have a separate replication data set.