1 Multiple Comparisons

Common questions:

1) After a significant ANOVA, which groups differ? (*Post hoc* tests)

2) While planning an experiment, you wish to test certain hypotheses that are a subset of the global ANOVA H_0. (*A priori* tests)

With the cuckoo data, the first question arises after finding that the ANOVA global $H_0 : \mu_1 = \mu_2 = ... = \mu_6$ is rejected, and wishing to investigate further which species differ. The second question could arise if while planning the study, we were specifically interested, for example, in whether mean egg length differed between Meadow Pipit and Tree Pipit nests. The problem that occurs in either of the two situations above is that when many individual tests are made, the chance of making a Type I error for at least one test can be far greater than the stated α level per test. For example, if 5 tests are made, each using $\alpha = .05$, then the probability of making a Type I error for at least one of the five tests (assuming independent tests) is $1 - (.95)^5 \approx .22$, which is far larger than .05. Multiple comparison methods aim to control the Type I error rate for a whole set of tests.

1.1 Terminology

When describing multiple comparison tests, it is useful to introduce some terminology. The term **contrast** is used to describe a comparison of means. Specifically, a contrast is a linear combination of the population means,

$$L = \sum_{i=1}^{t} a_i \mu_i \text{ that also satisfies } \sum_{i=1}^{t} a_i = 0.$$

We require that the coefficients a_i sum to zero so that the comparison is meaningful (we would not be interested in $\mu_1 - 3\mu_2$ for example). For contrasts we are generally interested in testing the null hypothesis $H_0 : L = \sum_{i=1}^{t} a_i \mu_i = 0$, against the alternative hypothesis $H_A : L = \sum_{i=1}^{t} a_i \mu_i \neq 0$. As an example, if we wished to test whether the mean cuckoo egg length from Meadow Pipit nests differed from that of Tree Pipits, we can express the null hypothesis as $H_0 : 1\mu_1 - 1\mu_2 = \mu_1 - \mu_2 = 0$. Here a_1=1 and a_2= -1, so a_1+ $a_2 = 0$ as required. This is an example of a **pairwise contrast**, which is defined as a contrast involving only two groups. An example of a **non-pairwise contrast** would be if we wished to test if the mean of cuckoo eggs in Pipit nests (Meadow and Tree) differed from cuckoo eggs in the other four groups. We can express this null hypothesis as $H_0 : (\mu_1 + \mu_2)/2 - (\mu_3 + \mu_4 + \mu_5 + \mu_6)/4 = 0$. Here a_1=1/2, a_2= 1/2, a_3= -1/4, a_4= -1/4, a_5= -1/4 and a_6= -1/4, so again a_1+ $a_2 + a_3 + a_4 + a_5 + a_6 = 0$, as required by the definition of a contrast. One other property of a set of contrasts, called orthogonality, is useful when considering *a priori* tests. Two contrasts

$$L_1 = \sum_{i=1}^{t} a_{1i} \mu_i \text{ and } L_2 = \sum_{i=1}^{t} a_{2i} \mu_i \text{ are } \textbf{orthogonal} \text{ if } \sum_{i=1}^{t} \frac{a_{1i} a_{2i}}{n_i} = 0.$$

If the group sample sizes are equal then this is equivalent to $\sum a_{1i} a_{2i} = 0$. In the examples above, if we identify the Meadow Pipit versus Tree Pipit contrast as L_1 and the Pipit versus other species contrast as L_2, then a_{11}=1, a_{12}= -1, $a_{13} = a_{14} = a_{15} = a_{16} = 0$, are the coefficients for L_1 and a_{21}=1/2, a_{22}= 1/2, a_{23}= -1/4, a_{24}= -1/4, a_{25}= -1/4 and a_{26}= -1/4 are the coefficients for L_2. Then if the sample sizes are equal,

$$\sum_{i=1}^{t} a_{1i}a_{2i} = (1)(1/2) + (-1)(1/2) + (0)(-1/4) + (0)(-1/4) + (0)(-1/4) + (0)(-1/4) = 0,$$

so L_1 and L_2 are orthogonal. Orthogonal contrasts are statistically independent, so that the outcome of testing one contrast is independent of the outcome of testing the other contrast. In our example, whether or not Meadow and Tree Pipits differ from each other (in terms of mean cuckoo egg length) gives no information about whether the average of Meadow and Tree Pipits differ from the average of the other four species. A set of more than two contrasts is **mutually orthogonal** if each pair of contrasts in the set is orthogonal to each other. The concept of a contrast or a set of contrasts at first seems somewhat esoteric, but in fact it is essential to understand these concepts to fully understand ANOVA, particularly in complicated situations.

2 A simple guide for multiple comparisons

There are a vast number of methods used for multiple comparison tests, and we will only consider a small number of them. We will only consider in detail three types of multiple comparison tests: **t tests**, **Tukey**'s method, and **Scheffe**'s method. **Fisher's LSD** method will also be discussed since it is widely used. The choice between these methods is governed by the type of contrasts being tested. In the somewhat artificial case in which a set of orthogonal contrasts has been specified *a priori*, then since the tests are independent we can simply apply separate t tests for each contrast, without adjusting the α level per contrast. If a set of contrasts has been specified *a priori* but are not orthogonal, t tests are again used but with a **Bonferroni correction**. In this case if c tests are involved, and the overall Type I error rate is to be held at α, then the significance level for individual tests is set at $\alpha' = \alpha/c$. Thus if 5 tests will be performed and the overall significance level for the set of tests is desired to be $\alpha = .05$, then $\alpha' = .05/5 = .01$ will be used for each individual test. If the contrasts to be tested are decided after collecting the data (*post hoc*) then we use generally more conservative methods to guard against data-snooping. For pairwise contrasts we can use Tukey's method and for non-pairwise contrasts we use Scheffe's method. This overall strategy is summarized in the following table, where the rows identify whether the contrasts are *a priori* or *post hoc*, and the columns identify whether they are orthogonal or not. Notice that all *post hoc* contrasts are treated as if they are nonorthogonal.

	Orthogonal	Nonorthogonal
A priori	Separate t-tests	Separate t-tests with Bonferroni correction
Post hoc		Pairwise: Tukey; Non-pairwise: Scheffe

3 Methods

3.1 t tests

We can estimate the contrast

$$L = \sum_{i=1}^{t} a_i\mu_i \text{ with } \widehat{L} = \sum_{i=1}^{t} a_i\overline{y}_{i.} \text{ and } \widehat{Var}(\widehat{L}) = \text{ MSE } \sum_{i=1}^{t} \frac{a_i^2}{n_i},$$

which leads to a t test of $H_0 : L = \sum_{i=1}^{t} a_i\mu_i = 0$,

$$t = \frac{\widehat{L}}{s.e.(\widehat{L})} = \frac{\sum_{i=1}^{t} a_i \overline{y}_{i.}}{\sqrt{\widehat{Var}(\widehat{L})}}.$$

For a two-tailed test, the t value is compared to $t_{df,\alpha/2}$, where df is the degrees of freedom for MSE ($df = t(n-1)$ for balanced 1 way ANOVA). Confidence intervals for L can also be constructed as $\widehat{L} \pm t_{df,\alpha/2}\ s.e.(\widehat{L})$.

3.2 Tukey's method for pairwise contrasts

Tukey's method is used for testing all pairwise differences between groups. Here we assume that the sample size is equal in each group, which is represented by n. Modifications are available when sample sizes differ between groups. To perform Tukey's method, follow these steps:

1. Rank the sample means.

2. Calculate $W = q_{t,df,1-\alpha}\sqrt{\frac{\text{MSE}}{n}}$, where $q_{t,df,1-\alpha}$ is the 100(1-α)% point of the Studentized range distribution, t is the number of groups, df is the degrees of freedom for MSE, and n is the common sample size per group.

3. Two population means μ_i and $\mu_{i'}$ are declared different if $|\overline{y}_{i.} - \overline{y}_{i'.}| \geq W$.

4. The results for the set of groups are often depicted graphically by drawing the ordered means and connecting groups that do not differ by a line.

Confidence intervals for μ_i - $\mu_{i'}$ can be constructed as $(\overline{y}_{i.} - \overline{y}_{i'.}) \pm W$. The Studentized range distribution used by Tukey is the distribution of the difference $\overline{y}_{MAX} - \overline{y}_{MIN}$ for a given number (t) of groups. In effect it is treating all pairwise comparisons as if they came from data snooping to pick the most different pair of means. Thus it controls the Type I error rate at α for the entire collection of pairwise tests. It cannot disagree with the result of the global ANOVA test, in the sense that if any pair of means are declared different by Tukey, then the global ANOVA H_0 must have been rejected.

3.3 Scheffe's method for general contrasts

For a general contrast $L = \sum_{i=1}^{t} a_i \mu_i$ Scheffe's method can be used to either test a hypothesis about L or construct a confidence interval. To test $H_0 : L = 0$, against $H_A : L \neq 0$ with Scheffe's method, follow these steps:

1. Calculate $\widehat{L} = \sum_{i=1}^{t} a_i \overline{y}_{i.}$.

2. Calculate $S = \sqrt{(t-1)F_{t-1,df,1-\alpha}}\sqrt{\widehat{Var}(\widehat{L})}$, where df is the degrees of freedom for MSE, and $\widehat{Var}(\widehat{L})$ is listed above.

3. If $|\widehat{L}| > S$ then reject H_0.

Confidence intervals for L can be constructed as $\widehat{L} \pm S$. Scheffe's method controls the Type I error rate at α for the entire collection of general contrasts, whether pairwise or nonpairwise. Like Tukey's method it cannot disagree with the result of the global ANOVA H_0. Scheffe's method is considered to be too conservative for use with pairwise comparisons.

3.4 Fisher's LSD method for pairwise contrasts

Another popular method for pairwise comparisons is Fisher's LSD (Least significant difference) method. In this method the means are ranked and then differences $(\overline{y}_{i.} - \overline{y}_{i'.})$ are compared to $LSD = t_{df,\alpha/2}\sqrt{\text{MSE}(\frac{1}{n_i} + \frac{1}{n_{i'}})}$. The text incorrectly describes the LSD method as using a Bonferroni correction when in fact the it is not used in practice. Many researchers advocate only performing Fisher's LSD method after rejecting the global ANOVA H_0, in which case it is described as Fisher's protected LSD method. Notice that if a pairwise contrast is being used, then Scheffe's S value reduces to $S = \sqrt{(t-1)F_{t-1,df,1-\alpha}}\sqrt{\text{MSE}(\frac{1}{n_i} + \frac{1}{n_{i'}})}$. A comparison of the LSD and S terms shows that Fisher's method uses $t_{df,\alpha/2} = \sqrt{t^2_{df,\alpha/2}} = \sqrt{F_{1,df,1-\alpha}}$, which differs from Scheffe's method in two ways: by not having a $(t-1)$ multiplier next to the F value, and by having only 1 numerator degree of freedom instead of $t-1$. Fisher's LSD method is therefore much more liberal than Scheffe's method, and in fact does not always control the Type I error rate at α even when only the protected method is used.

4 A final note

As previously stated, there are a vast number of multiple comparison methods in use. We have only discussed a very small number of methods that are widely recognized, implemented in most software, and all provide confidence intervals as well as hypothesis tests. There are, for example, methods for pairwise contrasts that control Type I error for the collection of tests as Tukey's method does, but have much greater power for detecting differences. A good discussion of many methods is found in Chapter 4 of Kirk (1995).

5 Reference

Kirk, R. E. (1995) Experimental Design: Procedures for the Behavioral Sciences. Pacific Grove: Brooks/Cole.