Multiple-Comparison Procedures

References

A good review of many methods for both parametric and nonparametric multiple comparisons, planned and unplanned, and with some discussion of the philosophical as well as practical issues in their use, is:

> Day, R. W., and G. P. Quinn (1989) “Comparisons of treatments after an analysis of variance in ecology,” *Ecological Monographs* **59**: 433-463.

Also, the GLM chapter of the SAS manual discusses the various error rates controlled by different procedures, gives formulas, and concludes with reasonable recommendations.

For a contrary perspective on the value of controlling the “overall significance level” for a set of inferences (and interesting discussion of a couple of other issues in statistics), see:

> Stewart-Oaten, A. (1995) “Rules and judgements in statistics: Three examples,” *Ecology* **76**: 2001-2009.

General usage

The purpose of most multiple-comparisons procedures is to control the “**overall significance level**” for some set of inferences performed as a follow-up to ANOVA. This “overall significance level” or error rate is the **probability, conditional on all the null hypotheses being tested being true, of rejecting at least one of them**, or equivalently, of having at least one confidence interval not include the true value.

Overview of methods

There are many methods for multiple-comparisons. Most are for pairwise comparisons of group means, to determine which are significantly different from which others. Other methods, though, are for more specialized purposes (*e.g.* comparing each of several treatments to a control) or allow testing of more-general hypotheses contrasting sets of group means (as is often done in preplanned “contrasts”).

The various methods differ in how well they properly control the overall significance level and in their relative power; some, such as the popular “Duncan’s multiple range test” **do not** control the overall significance level. The ones described in this handout all adequately control overall significance and are either easy to use or powerful. They are:

- Bonferroni   extremely general and simple, but often not powerful
- Tukey’s   the best for all-possible pairwise comparisons when sample sizes are unequal or confidence intervals are needed; very good even with equal samples sizes without confidence intervals
- stepdown   the most powerful for all possible pairwise comparisons when sample sizes are equal
- Dunnett’s   for comparing one sample (“control”) to each of the others, but not comparing the others to each other.

- MCB compares each mean to the "best" (largest or smallest, as you specify) of the other means.
- Scheffé's for unplanned contrasts among sets of means

When more than one method is applicable, and it is not clear which is more efficient (*i.e.* will give the narrower CI, for given α), it is legitimate to try all the applicable methods and use the one which proves most efficient for the given data.

Notation

In the following

- α = the "overall significance level" for the set of inferences,
- α^* = the significance level for a single one of those inferences,
- c = the number of inferences in the set,
- I = the number of samples, and
- N = the total sample size.

Note that when all possible pairwise comparisons are made among I means, there are $c = I(I\text{-}1)/2$ comparisons.

Bonferroni

This approach can be thought of as "alpha-splitting." If c inferences (tests or confidence intervals) are each made at some level α^*, the maximum possible "overall error rate" is $c\alpha^*$. Therefore you simply set the level for each separate inference, α^*, equal to α/c, where α is your desired overall significance level. Equivalently, and perhaps preferably, when the inferences are tests of hypotheses you can compute a *P*-value as usual for each test, and multiply it by c.

Note that **this method is completely general: it applies to *any* set of *c* inferences**, not only to multiple comparisons following ANOVA.

Advantages and disadvantages

The main advantage of this approach is that it is very easy, as well as very widely applicable. The main disadvantage is that it often is unnecessarily conservative (weak): α^* is smaller than it needs to be.

Procedures

Computer packages will apply these procedures to all possible pairwise comparisons; if the number of means is I, the total number of possible pairwise comparisons is $c = I(I\text{-}1)/2$. If you only want to compare some of the means, you can do so and define c accordingly. By doing this you do not pay as much of a multiple-comparisons penalty as if you did all possible comparisons. However, you must have decided which subset of comparisons to make before seeing the data!! Otherwise you implicitly did compare all means and must adjust the analysis accordingly.

If your software does not do Bonferroni comparisons, or if you want to coerce the software to correct for only a subset of all pairwise comparisons (as discussed in the preceding paragraph), you may be able to use “Fisher’s LCD” comparisons; these are pairwise comparisons with no correction to control the overall significance level. You therefore would specify the appropriate α^* for each comparison, based on the number of comparisons.

If doing Bonferroni comparisons by hand you can calculate either two-sample *t* statistics (or CIs) for each pair of means, or pairwise contrasts between means; the latter approach uses the ANOVA MSE and thus shares the ANOVA assumption of equal variances, while the former approach does not require this assumption. You then use the *t* critical values for $\alpha^*/2$ in the tests or CIs. (Since $\alpha^*/2$ probably will not appear in a table of *t* critical values, you may need to interpolate if doing this by hand).

Tukey’s (“honestly significant difference” or “HSD”)

This approach is specifically for comparing group means in an ANOVA setting. It is based on the distribution of *q*, the “studentized range.” The “studentized range” with *k* and *r* degrees of freedom is the range (*i.e.* maximum – minimum) of a set of *k* independent observations from some normal distribution, divided by an independent estimate (with *r* degrees-of-freedom) of the standard deviation of that normal distribution. Many texts have tables of this distribution.

If there are *I* samples, all populations’ means are the same (the complete null hypothesis is true), $\bar{x}_L$ and $\bar{x}_S$ are the largest and smallest sample means, and n_L and n_S are the respective sample sizes, then

$$(\bar{x}_L - \bar{x}_S) \Big/ \sqrt{\left(\frac{\text{MSE}}{2}\right)\left(\frac{1}{n_L} + \frac{1}{n_S}\right)}$$

will follow the studentized-range distribution with *I* and *N* - *I* degrees of freedom (*N* is the total sample size). Critical values for *q* then would be appropriate for comparing $\bar{x}_L$ and $\bar{x}_S$. Although other pairs of means do not actually represent the range of the observed sample of means (they will differ by less than $\bar{x}_L$ - $\bar{x}_S$), *q* critical values also are used for comparing them; this results in a conservative procedure.

A (1-α) CI for (μ_i - μ_j) is

$$(\bar{x}_i - \bar{x}_j) \pm q_{I,\, N-I,\, 1-\alpha} \sqrt{\left(\frac{\text{MSE}}{2}\right)\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$$

μ_i and μ_j are significantly different at level α if

$$\left|\bar{x}_i - \bar{x}_j\right| \geq q_{I,\, N-I,\, 1-\alpha} \sqrt{\left(\frac{\text{MSE}}{2}\right)\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$$

Advantages and disadvantages

When doing all pairwise comparisons, this method is considered the **best available when confidence intervals are needed or sample sizes are not equal**. When samples sizes are equal and confidence intervals are not needed Tukey's test is slightly less powerful than the stepdown procedures, but if they are not available Tukey's is the next-best choice, and unless the number of groups is large, the loss in power will be slight.

Procedures

Any worthwhile statistical software will provide Tukey's comparisons.

"Stepdown" procedures

As noted above, Tukey's procedure treats every pair of sample means as if they are the most-different pair: every pairwise difference is compared to the distribution expected for the range of *I* means. The stepdown procedures are modifications of Tukey's procedure which take into account that all but one of the comparisons are less different than the range. In essence, they work like this:

1. Compare the largest and smallest sample means, $\bar{x}_L$ and $\bar{x}_S$, to the *q* distribution for comparing *I* means. If this comparison is not significant, stop; otherwise, continue.
2. Compare the next most different pair of means (either the largest and the second-smallest or the second-largest and the smallest) to the *q* distribution for comparing *I* - 1 means. This is appropriate since one of the two most-extreme means is excluded from the set of means of which this second comparison represents the range. If this comparison is not significant, stop; otherwise, continue.
3. Continue comparing successively less-different pairs of means to *q* distributions with the "number of means" parameter successively smaller to represent the smaller set of means being considered in choosing each pair.
4. Etc. until a comparison is not significant or all comparisons have been made.

Advantages and disadvantages

When doing all pairwise comparisons, this approach is the **best available** when confidence intervals are not needed and sample sizes are equal.

Procedures

There actually are numerous implementations of the general step-down procedure described above. Which — if any — are available will depend on the software being used. SAS provides the "Ryan-Einot-Gabriel-Welsch multiple-range test" in the one-way ANOVA procedure within the "Analyst" application and presumably also within the "Enterprise Guide" interface (also available as the "REGWQ" and "REGWF" options to PROC GLM).

Dunnett’s procedure for comparing treatments with a control

When you do not wish to make all pairwise comparisons, but rather only to compare one of the groups (the “controls”, usually) with each other group, this procedure reduces the multiple-comparisons price you otherwise would pay, while taking into account that all these comparisons are correlated since they all use the same “control” data. You can also specify one-sided tests, which might be appropriate here but are not in general multiple comparisons. Unfortunately, the table of the critical values needed for this procedure is not widely available, so you probably will have to rely on a computer package.

Advantages and disadvantages

When it is appropriate — when you really are interested only in comparisons of one group to each of the others — this approach is **more powerful** than methods performing all possible pairwise comparisons, and therefore is recommended. Its disadvantage is simply that it does not compare the “other” groups to each other at all.

Hsu’s multiple comparisons with best (MCB)

“Hsu’s MCB” does comparisons between each sample mean and the “best” of all the other means, where you specify that “best” means either largest or smallest. In essence it is a modification of Dunnett’s method, allowing it to be applied when you do not know in advance which group you want to compare all the others to. Its purpose is as the name suggests: to select which group(s) is/are the best: not significantly different from each other but significantly better than the others.

Advantages and disadvantages

Since this approach does fewer comparisons than do methods (Tukey’s, stepdown) designed for doing all pairwise comparisons, it does not have to make as great a reduction in individual significance levels and therefore will be more powerful. Conversely, it makes more comparisons than Dunnett’s and therefore will be less powerful. It thus is the best method available when it truly is appropriate to your purpose.

Scheffé’s method for all possible contrasts

If you want to consider post hoc comparisons other than just pairwise comparisons (i.e. unplanned contrasts), Scheffé’s method can be used to control the overall confidence level. The best way to use this procedure is to calculate confidence intervals for the contrasts of interest, just as for preplanned contrasts but using (in place of the *t** critical value) the quantity *S*, where

$$S^2 = (I-1) \cdot F^*_{1-\alpha;I-1;N-I}$$

i.e. use (*I*-1) times the square root of the appropriate *F* critical value.

Advantages and disadvantages

This procedure is **extremely conservative** since it controls the overall significance for any possible contrast or set of contrasts, even when suggested by the data. It therefore **should not be used if only pairwise comparisons are wanted**. Furthermore, if the number of contrasts being considered is not considerably greater than the number of groups (i.e. *I*) and the contrasts were not suggested by the data, a **Bonferroni** correction probably will be more powerful than Scheffé's. **If the contrasts were suggested by the data**, however, **Scheffé's should be used** rather than Bonferroni, since all possible contrasts were implicitly considered (or at least you were willing to consider them).

Computing procedures

Minitab

Minitab provides only Tukey's, Dunnett's and Hsu's MCB comparisons; it does not provide any step-down methods.

Multiple comparisons procedures are options of the oneway ANOVA procedures (**Stat → ANOVA → Oneway …** and **Stat → ANOVA → Oneway (Unstacked)…**), as well as for some of the multifactor ANOVA procedures.

In the "One-way Analysis of Variance" window, after selecting the response variable and the factor, click on the **Comparisons …** button. Then select the procedure(s) desired; choices are Tukey's, Fisher's, Dunnett's, and Hsu's. Tukey's, Dunnett's, and Hsu's procedures are as described above, and each allows the "family" (*i.e.* overall) error rate to be specified. For Dunnett's, which level of the factor is to be considered the "control" must be specified (if values are text rather than numerical, the "control" value must be put in quotes).

The "Fisher's" option simply does pairwise comparisons with no attempt to control the overall significance level. If done only when the main ANOVA is significant, this is called "Fisher's protected Least Significant Difference (LSD)" procedure ("protected" by the significance of the main ANOVA). As noted in the section on the Bonferroni method, this procedure also can be used to carry out Bonferroni comparisons by specifying the appropriate <u>individual</u> error rate (*i.e.* $\alpha^* = \alpha/c$ where α is the desired overall significance and c is the total number of comparisons desired); this would only be useful if for some reason only a subset (chosen in advance) of the pairwise comparisons was of interest so that Tukey's procedure would be too conservative.

<u>Results:</u>

Results are presented as confidence intervals for pairwise differences. After a heading saying what kind of comparisons they are and at that confidence level, there is a line stating the "test-wise" confidence level corresponding to the chosen "family-wise" confidence level.

Then come the actual results, in the form of a series of tables, each comparing one group to each of the subsequent groups. The first table compares the first group to each of the others, the second table compares the second group to all remaining groups (i.e. all but the first), etc. (Group ordering is alphabetical by the group labels.) Each table is labelled "`*factor* = *value* subtracted from:`" where the name of the grouping variable (the ANOVA factor) takes the place of *`factor`* and the label of the reference group for that table takes the place of *`value`*. The rows in the table then are labelled (in the column headed "*`factor`*") by the label of the group being compared to that table's reference group. The pairwise comparisons in each row are given as

- the estimated difference in population means (estimated simply by the difference in sample means), listed in the column headed "`Center`", surrounded by
- the CI for the difference in population means, listed in the columns headed "`Lower`" and "`Upper`", followed by
- a crude chart of these estimates.

```
Tukey 95% Simultaneous Confidence Intervals
All Pairwise Comparisons among Levels of source

Individual confidence level = 98.06%


source = Kauai subtracted from:

source    Lower   Center   Upper   --+---------+---------+---------+-------
kure    -1.3967  -0.7873  -0.1780     (--------*-------)
midway  -1.5076  -0.5698   0.3680   (-------------*------------)
                                   --+---------+---------+---------+-------
                                   -1.40     -0.70      0.00      0.70


source = kure subtracted from:

source    Lower  Center   Upper   --+---------+---------+---------+-------
midway  -0.7906  0.2175  1.2256              (-------------*--------------)
                                  --+---------+---------+---------+-------
                                 -1.40     -0.70      0.00      0.70
```

S-Plus

S-Plus provides several methods, including Tukey’s, Dunnett’s, and a special simulation-based method. It does not provide Hsu’s MCB or any step-down methods.

Multiple comparisons can be done using the general ANOVA procedure invoked by **Statistics → ANOVA → Fixed Effects…** (The simpler method invoked by Statistics → Compare Samples → k Samples → Oneway ANOVA … does not do comparisons.) Comparisons are requested on the **Compare** tab of the ANOVA dialog. The only item in this dialog that must be specified is the **Variable - Levels Of:**, which must be the factor in the ANOVA model. The default options generally are appropriate.

Options:

The **Comparison Type** part of the dialog gives a choice among

- **mca** (mean comparisons among all groups; this is the default),
- **mcc** (mean comparisons with control), and
- **none**, which gives CIs for the level means rather than pairwise comparisons.

The **Method** part of the dialog gives a choice among **best**, **best.fast**, **Bonferroni**, **Dunnett**, **Fisher.lsd**, **Scheffe**, **Sidak**, **Simulation**, and **Tukey**. Most of these are described in the main part of this handout. The others are:

- **best**: whichever of the suite of methods is most powerful,
- **best.fast**: whichever of the suite of methods other than Simulation is most powerful,
- **Fisher.lsd**: as described in passing elsewhere in this handout, these are pairwise contrasts with no correction for multiple testing,
- **Sidak**: similar to but slightly more powerful than Bonferroni, and
- **Simulation** (see below).

The default method is **best.fast**. When the “Comparison Type” is “mca” the “best.fast” option generally will use the Tukey method for two-ended CIs and the Sidak method if only upper or lower bounds are requested. When the “Comparison Type” is “mcc” Dunnett’s method generally will be used.

The **Simulation** method uses “Monte Carlo” simulation of the sampling distribution of the set of sample means assuming the null hypothesis of equal population means. This distribution will depend on will details of the design (including sample sizes) in ways that the standard methods, using tabled distributions, cannot fully address. The simulation method therefore will generally be more powerful than other valid methods, and sometimes will be much more powerful. Its disadvantage is that it can take a lot of computer time with large data sets.

Output:

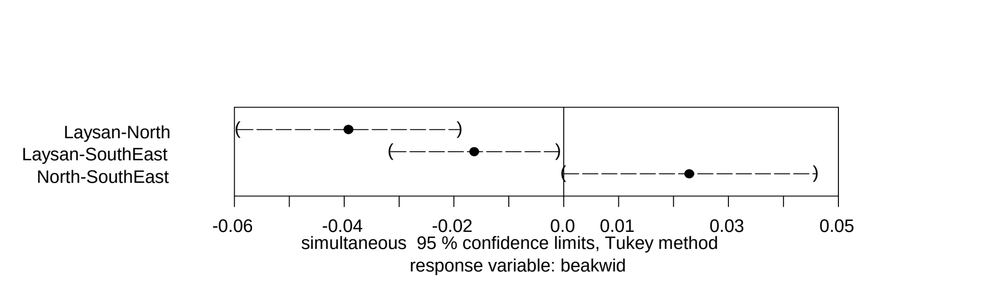
The output (see below) is in the form of the difference, its SE, and the resulting CI, for each of the pairs of differences. Significant comparisons are flagged. If the **Plot Intervals** option is checked, a graph showing all the pairwise CIs is produced, as shown below after the text output.

```
95 % simultaneous confidence intervals for specified
linear combinations, by the Tukey method

critical point: 2.3836
response variable: beakwid

intervals excluding 0 are flagged by '****'

                 Estimate Std.Error Lower Bound Upper Bound
    Laysan-North  -0.0392   0.00848  -0.0594000    -0.01900 ****
Laysan-SouthEast  -0.0163   0.00640  -0.0315000    -0.00104 ****
 North-SouthEast   0.0229   0.00964  -0.0000741     0.04590
```


Laysan-North
Laysan-SouthEast
North-SouthEast
simultaneous 95 % confidence limits, Tukey method
response variable: beakwid