Comparing Multiple Comparisons

Phil Ender

Culver City, California

Stata Conference Chicago - July 29, 2016

Prologue

In ANOVA, a significant omnibus F-tests only indicates that there is a significant effect.

It does not indicate where the significant effects can be found.

This is why many, if not most, significant ANOVAs, with more than two levels, are followed by post-hoc multiple comparisons.

What's is the Problem?

Computing multiple comparisons increases the probability of making a Type I error.

The more comparisons you make, the greater the chance of Type I errors.

Multiple comparison techniques are designed to control the probability of these Type I errors.

What's the Problem? Part 2

If n independent contrasts are each tested at α, then the probability of making at least one Type I error is $1 - (1 - \alpha)^n$.

The table below gives the probability of making at least one type I error for different numbers of comparisons when $\alpha = 0.05$:

n	probability
1	0.0500
2	0.0975
3	0.1426
5	0.2262
10	0.4013
15	0.5367
20	0.6415

The above probabilities apply to independent contrasts. However, most sets of contrasts are not independent.

What is the solution?

Adjust the critical values or p-values to reduce the probability of a false positive.

The goal is to protect the familywise or experimentwise error rate in a strong sense, i.e., whether the null is true or not.

Multiple comparison techniques such as Dunnett, Tukey HSD, Bonferroni, Šidàk or Scheffè do a reasonably good job of of protecting the familywise error rate.

Techniques such as Fisher's least significant difference (LSD), Student-Newman-Keuls, and Duncan's multiple range test fail to strongly protect the familywise error rate. Such procedures are said to protect the familywise error rate in a weak sense, avoid them if possible.

Outline of Multiple comparisons

I. Planned Comparisons

A. Planned Orthogonal Comparisons

B. Planned Non-orthogonal Comparisons

II. Post-hoc Comparisons

A. All Pairwise

B. Pairwise versus control group

C. Non-pairwise Comparisons

III. Other Comparisons

I. Planned Comparisons

Planned Orthogonal Comparisons

These are among the most powerful hypothesis tests available.

- Two Stringent requirements:

Planned Orthogonal Comparisons

These are among the most powerful hypothesis tests available.

- Two Stringent requirements:
- 1. Comparisons must be planned

Planned Orthogonal Comparisons

These are among the most powerful hypothesis tests available.

- Two Stringent requirements:
- 1. Comparisons must be planned
- 2. Comparisons must be orthogonal

Planned Orthogonal Comparisons

These are among the most powerful hypothesis tests available.

- Two Stringent requirements:
- 1. Comparisons must be planned
- 2. Comparisons must be orthogonal

- Say, 1vs2, 3vs4 and avg 1&2vs avg 3&4

Planned Orthogonal Comparisons

These are among the most powerful hypothesis tests available.

- Two Stringent requirements:
- 1. Comparisons must be planned
- 2. Comparisons must be orthogonal

- Say, 1vs2, 3vs4 and avg 1&2vs avg 3&4

- Downside: Comparisons of interest may not be orthogonal.

Planned Non-orthogonal Comparisons

Use either the Dunn or the Šidàk-Dunn adjustment.

Consider C contrasts:

Dunn: $\alpha_{Dunn} = \alpha_{EW}/C$

Šidàk-Dunn: $\alpha_{SD} = 1 - (1 - \alpha_{EW})^{(1/C)}$

If $C = 5$ and $\alpha_{EW} = .05$ then $\alpha_{Dunn} = .01$ and $\alpha_{SD} = .010206$.
Basically, just Bonferroni and Šidàk adjustments.

Planned Non-orthogonal Comparisons: Pairwise vs Control

Special Case: Pairwise versus control group.

Dunnett's test is used to compare $k - 1$ treatment groups with a control group. Does not require an omnibus F-test.

Dunnett's test is a t-test with critical values derived by Dunnett (1955). The critical value depends on the number of groups and the denominator degrees of freedom.

II. Post-hoc Comparisons

Post-hoc Comparisons: All pairwise

Tukey's HSD (honestly significant difference) is the perennial favorite for performing all possible pairwise comparisons among group means.

With k groups there are $k * (k - 1)/2$ possible contrasts.

Tukey's HSD uses quantiles of Studentized Range Statistic to make adjustments for the number of comparisons.

All pairwise contrasts with large k may look like a fishing expedition.

Post-hoc Comparisons: All pairwise

Tukey HSD Test,

$$q_{HSD} = \frac{Y_{mi} - Y_{mj}}{\sqrt{MS_{error}/n}}$$

Note the single *n* in the denominator. Tukey's HSD requires that all groups must have the same number of observations.

What if the cell sizes are not equal?

Harmonic mean, the old school approach

$$n = k/(1/n1 + 1/n2 + 1/n3 + 1/n4)$$

Spjøtvol and Stoline's modification of the HSD test,

$$q_{SS} = \frac{Y_{mi} - Y_{mj}}{\sqrt{MS_{error}/n_{min}}}$$

Uses the minimum *n* of the two groups. Uses Studentized Augmented Range distribution for *k* and error df.

More on unequal cell sizes

Tukey-Kramer Modification of the HSD test,

$$q_{TK} = \frac{Y_{mi} - Y_{mj}}{\sqrt{MS_{error}(1/n_i + 1/n_j)/2}}$$

Use the Studentized Range distribution for k means with ν error degrees of freedom.

Post-hoc Comparisons: Pairwise vs Control

I know Dunnett's test is for planned comparisons of $k-1$ treatment groups with a control group. However, it is also used for post-hoc comparisons. It is marginally more powerful then the Tukey HSD because there are fewer contrasts.

Dunnett's test is a t-test with critical values derived by Dunnett (1955). The critical value depends on number of groups (k) and the anova error degrees of freedom.

Post-hoc Comparisons: Non-pairwise Comparisons

Example: Average of groups 1 & 2 versus the mean of group 3.

Use the Scheffé adjustment.

Scheffé is very conservative adjustment making use the F *distribution. The Scheffé critical value is ...*

$F_{Crit} = (k - 1) * F_{(1,\nu error)}$

Where k is the total number of groups.

III. Other Comparisons

If you absolutely positively have to make a few comparisons, but ...

- but they don't fit any of the approaches we've seen so far?

If you absolutely positively have to make a few comparisons, but ...

- but they don't fit any of the approaches we've seen so far?
- ... say, 15 regressions on 15 separate response variables.

If you absolutely positively have to make a few comparisons, but ...

- but they don't fit any of the approaches we've seen so far?
- ... say, 15 regressions on 15 separate response variables.
- Try a Bonferroni or Šidák adjustments

If you absolutely positively have to make a few comparisons, but ...

- but they don't fit any of the approaches we've seen so far?
- ... say, 15 regressions on 15 separate response variables.
- Try a Bonferroni or Šidák adjustments
- Good protection but low power.

What if you want to make a huge number of contrasts, ...

- say 10,000 or more?

What if you want to make a huge number of contrasts, ...

- say 10,000 or more?
- Try a false discovery rate (FDR) method such as Benjamini-Hochberg.

What if you want to make a huge number of contrasts, ...

- say 10,000 or more?
- Try a false discovery rate (FDR) method such as Benjamini-Hochberg.
- FDR control offers a way to increase power while maintaining some principled bound on error.

What if you want to make a huge number of contrasts, ...

- say 10,000 or more?
- Try a false discovery rate (FDR) method such as Benjamini-Hochberg.
- FDR control offers a way to increase power while maintaining some principled bound on error.
- Note that when the FDR is controlled at .05, it is guaranteed that on average only 5% of the tests that are rejected are spurious.

What if you don't want to be bothered making any adjustments for multiple comparisons?

- Analyze your experiment using Bayesian methods.

What if you don't want to be bothered making any adjustments for multiple comparisons?

- Analyze your experiment using Bayesian methods.
- All comparisons are made from a single posterior distribution.

What if you don't want to be bothered making any adjustments for multiple comparisons?

- Analyze your experiment using Bayesian methods.
- All comparisons are made from a single posterior distribution.
- See whether the region of equivalence for the difference in means falls outside of the 95% highest posterior density (HPD) credible interval.

References

Benjamini, Y, & Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Statist Soc. Series B (Methodological), 57(1), 289.-300.

Hays, R.E. (1995). Experimental design: Procedures for the behavioral sciences (3rd Edition). Pacific Grove, CA: Brooks/Cole.

Kruschke, J.K. (2015). Doing bayesian analysis: a tutorial with R., JAGS and Stan (2nd Edition). Amsterdam: Elsevier.

¿Questions?