Post Hoc Tests

Familywise Error

Familywise error (FWE) is also known as alpha inflation or cumulative Type I error. Familywise error represents the probability that any one of a set of comparisons or significance tests is a Type I error. As more tests are conducted, the likelihood that one or more are significant just due to chance (Type I error) increases. One can estimate familywise error with the following formula:

$$\alpha_{FWE} \leq 1 - (1 - \alpha_{EC})^{K}$$

where α_{FWE} is the familywise error rate, α_{EC} is the alpha rate for an individual test (almost always considered to be .05), and K is the number of comparisons. K as used in the formula is an exponent, so the parenthetical value is raised to the K^{th} power.

Bonferroni

The Bonferroni (or sometimes referred to as the Dunn-Bonferroni) test is designed to control the familywise error rate by simply calculating a new pairwise alpha to keep the familywise alpha value at .05 (or another specified value). The formula for doing this is as follows:

$$EC = \alpha_B = \frac{\alpha_{FWE}}{K}$$

where EC is the new alpha based on the Bonferroni test that should be used to evaluate each comparison or significance test, α_{FWE} is the familywise error rate that is desired (often .05, but not necessarily), and K is the number of comparisons (statistical tests).

The Bonferroni is probably the most commonly used post hoc test, because it is highly flexible, very simple to compute, and can be used with any type of statistical test (e.g., correlations)—not just post hoc tests with ANOVA. The traditional Bonferroni, however, tends to lack power (Olejnik, Li, Supattathum, & Huberty, 1997). The loss of power (i.e., Type II errors are more likely), which is worse for more comparisons, occurs for several reasons: (1) the familywise error calculation depends on the assumption that, for all tests, the null hypothesis is true. This is unlikely to be the case, especially after a significant omnibus test; (2) all tests are assumed to be *orthogonal* (i.e., independent or nonoverlapping) when calculating the familywise error test, and this is usually not the case when all pairwise comparisons are made; (3) the test does not take into account whether the findings are consistent with theory and past research. If consistent with previous findings and theory, an individual result should be less likely to be a Type I error; and (4) with the Bonferroni correction, Type II error rates are too high for individual tests. In other words, then, the Bonferroni overcorrects for Type I error.

Modified Bonferroni Approaches

Several alternatives to the traditional Bonferroni have been developed, including those developed by Holm, Holland and Copenhaver, Hommel, Rom, and others (see Olejnik et al., 1997 for a review). These tests have greater power than the Bonferroni while retaining its flexible approach that allows for use with any set of statistical tests (e.g., t-tests, correlations, chi-squares).

Sidak-Bonferroni. Sidak (1967) suggested a relatively simple modification of the Bonferroni formula that would have slightly less of an impact on statistical power but retain much of the flexibility of the Bonferroni method (Keppel & Wickens, 2004, discuss this testing approach). Instead of dividing by the number of comparisons, there is a slightly more complicated formula:

$$\alpha_{S-B} = 1 - \left(1 - \alpha_{FWE}\right)^{1/K}$$

where $\alpha_{S\text{-}B}$ is the Sidak-Bonferroni alpha level used to determine significance (something less than .05), α_{FWE} is the desired familywise error (e.g., .05, or level desired by the researcher), and *K* is the number of comparisons or statistical tests conducted in the "family." The *p*-values obtained from the computer printout must be smaller than $\alpha_{S\text{-}B}$ to be considered significant. One can also extend this test to other statistical tests, such as correlations, and, therefore, it is a flexible adjustment. In the case of correlations, one could replace df_A with the number of variables that are used in the group of correlations tests. K would represent the number of correlations in the correlation matrix. This approach is convenient and easy to do but has not received much systematic study, and it is likely that a single, simple correction will not result in the most efficient balance of Type I and Type II errors.

Sequential Methods

Sequential tests involve a process that requires conducting pairwise comparisons and then ordering the *p*-values, where each subsequent decision for significance is dependent on the prior significance decision(s). Hochberg's sequential method (Hochberg, 1988; and Holm, 1979, proposed a similar step-down method)[1] is a "step-up" approach as a more powerful alternative to the Bonferroni procedure. Sequential methods use a series of steps in the correction, depending on the result of each prior step. Contrasts are initially conducted and then ordered according to *p*-values (from smallest to largest in the "step-up" approach). Each step corrects for the previous number of tests rather than all the tests in the set. This test is a good, high-powered alternative to the other modified Bonferroni approaches as long as confidence intervals are not needed. Unfortunately, this approach is not available in some statistical packages, like SPSS, but there is a spreadsheet method available online, http://www.real-statistics.com/hypothesis-testing/familywise-error/holms-and-hochbergs-tests/. The `p.adjust` function in the `stats` package in R conducts Holm, Hochberg, Hommel, and Benjamin-Hochberg (for false discovery—see below) tests.

Approaches for Pairwise Comparisons with ANOVA Designs

Dunn. Identical to the Bonferroni correction.

Scheffe. The Scheffe test computes a new critical value for an F test conducted when comparing two groups from the larger ANOVA (i.e., a correction for a standard t-test). The formula simply modifies the F-critical value by taking into account the number of groups being compared: $(a-1)\,F_{crit}$. The new critical value represents the critical value for the maximum possible familywise error rate. As you might suppose, this also results in a higher than desired Type II error rate, by imposing a severe correction.

Fisher LSD. The Fisher LSD test stands for the Least Significant Difference test (rather than what you might have guessed). The LSD test is simply the rationale that if an omnibus test is conducted and is significant, the null hypothesis is *incorrect*. (If the omnibus test is nonsignificant, no post hoc tests are conducted.) The reasoning is based on the assumption that if the null hypothesis is incorrect, as indicated by a significant omnibus F-test, Type I errors are not really possible (or less likely), because they only occur when the null is true. So, by conducting an omnibus test first, one is screening out group differences that exist due to sampling error, and thus reducing the likelihood that a Type I error is present among the means. Fishers LSD test has been criticized for not sufficiently controlling for Type I error. Still, the Fisher LSD is sometimes found in the literature.

Dunnet. The Dunnet test is similar to the Tukey test (described below) but is used only if a set of comparisons are being made to one particular group. For instance, we might have several treatment groups that are compared to one control group. Since this is rarely of interest, and the Tukey serves a much more general purpose, I recommend the Tukey test.

[1] This step-up procedure is not the same as what SPSS calls Hochberg's GT2, which is Hochberg's original proposed method of FWE control (Hochberg, 1974).

Tukey a (also known as Tukey's HSD for honest significant difference). Tukey's test calculates a new critical value that can be used to evaluate whether differences between any two pairs of means are significant. The critical value is a little different because it involves the mean difference that has to be exceeded to achieve significance. So one simply calculates one critcal value and then the difference between all possible pairs of means. Each difference is then compared to the Tukey critical value. If the difference is larger than the Tukey value, the comparison is significant. The formula for the critical value is as follows:

$$\overline{d}_T = q_T \sqrt{\frac{MS_{s/A}}{n}}$$

q_T is the studentized range statistic (similar to the t-critcal values, but different), which one finds in a table (Table C.9 in the Myers & Well text), $MS_{s/A}$ is the mean square error from the overall F-test, and n is the sample size for each group. *Error df* referred to in the table is the $df_{s/A}$ used in the ANOVA test. *FWE* is the desired familywise error rate. This is the test I usually recommend, because studies show it has greater power than the other tests under most circumstances and it is readily available in computer packages. The Tukey-Kramer test is used by SPSS when the group sizes are unequal. It is important to note that the power advantage of the Tukey test depends on the assumption that all possible pairwise comparisons are being made. Although this is usually what is desired when post hoc tests are conducted, in circumstances where not all possible comparisons are needed, other tests, such as the Dunnett or a modified Bonferroni method should be considered because they may have power advantages.[2]

Games-Howell. This test is used with variances are unequal (see Unequal Variances below) and also takes into account unequal group sizes. Severely unequal variances can lead to increased Type I error, and, with smaller sample sizes, more moderate differences in group variance can lead to increases in Type I error. The Games-Howell test, which is designed for unequal variances, is based on Welch's correction to *df* with the *t*-test and uses the studentized range statistic. This test appears to do better than the Tukey HSD if variances are very unequal (or moderately so in combination with small sample size) or can be used if the sample size per cell is very small (e.g., <6).

Comments

One difficulty that researchers often experience is a dilemma about what constitutes a family. There is no definite answer to this question. Keselman and colleagues (2011, p. 1) state, "A family of tests refers to a set of conceptually related hypotheses/tests; specification of a family of tests, self-defined by the researcher, can vary depending on the research paradigm." This remains a fairly ambiguous definition. Generally, most researchers consider all possible pairwise comparisons following an ANOVA as a family of multiple tests. But should a set of correlation coefficients be considered a family? Should multiple tests in an article or from a study be considered a family? Few researchers seem to consider the latter definition, but it raises the question of what the limit to a family should be.

Klockars, Hancock, and McAweeney (1995) discuss many of the post hoc ANOVA procedures, some of which seem to advantages over the traditional approaches such as the Tukey currently available in statistical software packages. Several distinctions among various tests can be made, including sequential vs. simultaneous, weighted vs. unweighted, and step-up vs. step-down, and they involve elaborate computational procedures which are inconvenient to do by hand especially for a large number of comparisons. Modified Bonferroni procedures have been designed for a broader array of statistical circumstances beyond post hoc ANOVA tests (e.g., correlations or chi-square tests). Olejnik and colleagues (1997) review the modified Bonferroni procedures and their computations. They conclude that most of the modified Bonferroni procedures have clear advantages over the traditional Bonferroni procedure, but small differences among the alternatives in the amount of power or control of Type I error. Their results suggest that Rom's (1990) procedure has the most power (not currently available in SPSS).

[2] To maximize power of the application of the Tukey test, one should still examine the comparisons even when the overall ANOVA is not significant. The requirement of the initial significance of the ANOVA tests to reduce its power performance (Ramsey & Ramsey, 2008).

To further complicate matters, Benjamini and Hochberg (1995) introduced an alternative conceptualization to familywise error called "false discovery rate." The false discover rate is the expected proportion of number of true null hypothesis rejections out of the total number null hypothesis rejections. The proposed method for controlling false discovery involves an ordering of p-values for all comparisons and then stepping down in significance decisions. The false discovery rate approach is more liberal than traditional familywise error control approaches, because it does not conceptualize the alpha inflation problem as a probability of making one or more Type I errors. Several proposed tests, which also do not attempt to strictly limited familywise error to just the probability of one or more Type I errors, set the criteria allowing for two or more (up to K-1 or K more) Type I errors above the chosen familywise error rate (Keselman et al, 2011). Together, false discovery and the K-more tests tend to be more powerful (and thus reducing Type II error) but at the potential cost of increasing Type I error.[3]

Other authors have reviewed post hoc tests with additional attention to unequal error variances (e.g., Kromrey & La Rocca, 1995; Seaman, Levin, & Serlin,1991). How heterogeneous (i.e., unequal) the error variances must be in order to cause problems is difficult to discern, because their impact is greater with lower sample sizes. Unfortunately, tests such as Levene's test for unequal variances have lower power when sample size is smaller, so they may be least likely to indicate a problem with unequal variances when it is most likely to affect Type I errors. In terms of post ANOVA tests, the Games-Howell is good if there are large differences in variances between groups.

I have included only a subset of all the possible post hoc corrections for familywise error. And, believe it or not, familywise error correction procedures currently available in most statistical packages (only some of which I have focused on here) represent only a subset of the approaches which have been proposed and studied. Many of the tests that appear to have the best Type I error control with the most power are not widely available in software packages. Among the tests available in SPSS (and several other packages) for ANOVA-design post hoc tests, the Tukey a (or "HSD" and Tukey-Kramer for unequal N and Games-Howell for unequal variances) is probably the most reasonable balance of power and Type I error control among the conventional tests available. If you want to maximize power and control Type I error then I suggest going to the trouble of conducting the Hochberg sequential test.

References

Benjamini, Y., & Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple hypothesis testing. *Journal of the Royal Statistical Society B, 57*:289–300

Hochberg, Y. (1974). Some generalizations of the T-method in simultaneous inference. *Journal of Multivariate Analysis, 4*, 224-234.

Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika, 75*, 800–802.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics, 6*, 65-70.

Keselman, H. J., Miller, C. W., & Holland, B. (2011). Many tests of significance: New methods for controlling type I errors. *Psychological Methods, 16*, 420.

Keppel, G., & Wickens, T.D. (2004). *Design and analysis: A researchers handbook* (4rd Edition). Upper Saddle River, NJ: Pearson.

Klockars, A.J.,Hancock, G.R., & McAweeney, M.J. (1995). Power of unweighted and weighted versions of simultaneous and sequential multiple-comparison procedures. *Psychological Bulletin, 118*, 300-307.

Kromrey, J.D., & La Rocca, M.A. (1995). Power and Type I error rates of new pairwise multiple comparison procedures under heterogeneous variances. Journal of Experimental Education, 63, 343-362.

McDonald, J.H. 2014. *Handbook of Biological Statistics, 3rd ed*. Sparky House Publishing, Baltimore, Maryland.

Olejnik,S., Li, J., Supattathum, S., and Huberty, C.J. (1997). Multiple testing and statistical power with modified Bonferroni procedures. *Journal of educational and behavioral statistics, 22*, 389-406.

Ramsey, P. H., & Ramsey, P. P. (2008). Power of pairwise comparisons in the equal variance and unequal sample size case. *British Journal of Mathematical and Statistical Psychology, 61*, 115-131.

Seaman, M.A., Levin, J.R., & Serlin, R.C. (1991). New developments in pairwise multiple comparisons: Some powerful and practicable procedures. *Psychological Bulletin, 110*, 577-586.

[3] See Keselman and colleagues also for R code for computing the Benjamini and Hochberg false discovery correction as well as several of these alternative K-more tests that allow two-or-more Type I errors above the *FWE* rate. MacDonald (2014) has posted an excel spreadsheet that will compute the Benjamini-Hochberg false discovery corrections, http://www.biostathandbook.com/benjaminihochberg.xls. See Charles Zaiontz's site for how to program an Excel sheet to do step-up and step-down tests, http://www.real-statistics.com/hypothesis-testing/familywise-error/holms-and-hochbergs-tests/, and, in SAS, PROC MULTITEST computes the Hochberg step-up and a number of other tests.