A Review of Experimentwise Type I Error: Implications for Univariate Post Hoc and for Multivariate Testing

Dan Altman

Texas A&M University 77843-4225

Paper presented at the annual meeting of the Southwest Educational Research Association, Dallas, January 29, 2000.

ABSTRACT

Experimentwise error rates can rapidly inflate when researchers use multiple univariate tests. Both (a) ANOVA post hoc and (b) multivariate methods incorporate a correction for experimentwise error. Researchers ought to understand experimentwise error if they are to understand (a) what post hoc test really do and (b) an important rationale for multivariate methods.

A REVIEW OF EXPERIMENTWISE TYPE I ERROR:
IMPLICATIONS FOR UNIVARIATE POST HOC AND FOR MUTIVARIATE TESTING

Researchers are wary of making a Type I error. In order to guard against doing that, researchers set alpha to be small. However, some researchers, focus only on “testwise” alpha, and are unaware of the “experimentwise” alpha and the iimportance of not inflating “experimentwise” Type I error rates. This paper reviews experimentwise Type I error. The concept is fundamentally important in two respects. First, ANOVA post hoc tests implicitly incorporate a correction for experiemtnwise error; if this correction is not understood, the researcher does not understand post hoc tests themselves. Second, experimentwise error concerns are one reason why multivariate tests are almost always vital in educational research (Fish, 1988; Thompson, 1999), so researchers ought to understand experimentwise error if they are to understand an important rationale for multivariate methods.

Experimentwise Error

Researchers are cognizant of the possibility of rejecting a null hypothesis (H_O) even when the H_O is true. This is called a "testwise" Type I error. Researchers set an alpha (α) level a priori at a small near-zero value to protect against testwise Type I errors. If the alpha level is set at .01 of statistical significance, one percent of the time the null will be falsely rejected. In this case, the null is rejected even though the null may be true in the population.

Most researchers are familiar with "testwise" alpha (α_{TW}). However, while "testwise" alpha refers to the probability of making a Type I error for a given hypothesis test, "experimentwise" (or "familywise" -- see Maxwell, 1992, p. 138) error rate refers to the probability of having made a Type I error anywhere within a set of hypothesis tests (Thompson, 1994). "Experimentwise" error rate inflates when a number of hypotheses are tested (e.g., two or more dependent variables) at the same alpha level within a given study (Love, 1988).

"Experimentwise" error rate equals "testwise" error rate when only one hypothesis is tested for a given group of people in a study. However, when more than one hypothesis is being tested in a given study with only one sample, the two error rates may not be equal (Thompson, 1994). This occurs as Type I errors from each individual tested hypothesis build off each other, causing a highly inflated experimentwise error rate. Huberty and Morris (1989, p. 306) referred to this as "probability pyramiding." Given the number of hypotheses being tested, the inflation of experimentwise error rates can be quite serious, as emphasized by Morrow and Frankiewicz (1979).

Experimentwise and testwise error rates are equal given the presence of multiple hypothesis tests (e.g., two or more dependent variables) in a single sample study only if the hypotheses (or the dependent variables) are perfectly correlated (or independent). This is so by reason that, for example, when one has perfectly correlated hypotheses, one actually is still only testing a single hypothesis. Therefore, it can be said that two factors effect the inflation of experimentwise Type I error: (a) the number of hypotheses tested using a single sample of data, and (b) the degree of correlation among the dependent variables or the hypotheses tested (Thompson, 1994).

Bonferroni Formula for α_{EW}

"Experimentwise" error rate inflation is at its maximum when multiple dependent variables (e.g., multiple hypothesis tests) in a single sample study are perfectly uncorrelated (Fish, 1988). When this occurs, the experimentwise error (α_{EW}) rate can be calculated. This is done using what is called the Bonferroni inequality (Love, 1988):

$$\alpha_{EW} = 1 - (1 - \alpha_{TW})^{K},$$

where k is the number of perfectly uncorrelated hypotheses or variablesbeing tested at a given testwise alpha level (α_{TW}).

For example, if four perfectly uncorrelated hypotheses (or dependent variables) are tested using data from a single sample, each at the α_{TW} = .01 level of statistical significance, the experimentwise Type I error rate will be:

$$\begin{aligned} \alpha_{EW} &= 1 - (1 - \alpha_{TW})^{K} \\ &= 1 - (1 - .01)^{4} \\ &= 1 - (\ .99\ )^{4} \\ &= 1 - (.99(.99)(.99)(.99)) \\ &= 1 - .960596 \\ \alpha_{EW} &= .039404. \end{aligned}$$

Thus, for a study testing four perfectly uncorrelated dependent variables, each at the α_{TW} = .01 level of statistical significance, the probability is .039404 (or 3.9404%) that one or more null hypotheses will be incorrectly rejected within the study. However, knowing this will not inform the researcher as to which one or more of the statistically significant hypotheses is a Type I error. Table 1 provides an illustration of these calculations for several α_{TW} levels. This table also illustrates how quickly α_{EW} can become inflated.

Witte (1985) explains the two error rates using an intuitively appealing example involving a coin toss. If the toss of heads is equated with a Type I error, and if a coin is tossed only once, then the probability of a head on the one toss (α_{TW}), and of at least one head within a set (α_{EW}) consisting of one toss, will both equal 50%.

If the coin is tossed three times, the "testwise" probability of a head on each toss is still 50%, i.e., α_{TW} = .50 (not .05). The Bonferroni inequality is a literal fit to this example situation (i.e., that is, a literal analogy), because the coin's behavior on each flip is literally uncorrelated with the coin's behavior on previous flips. In other words, the coin does not alter its behavior on any given flip as a result of its behavior on any previous flip.

Thus, the "experimentwise" probability (α_{EW}) that there will be at least one head in the whole set of three flips will be exactly:

$$
\begin{aligned}
\alpha_{EW} &= 1 - (1 - \alpha_{TW})^{K} \\
&= 1 - (1 - .50)^{3} \\
&= 1 - (\ .50\ )^{3} \\
&= 1 - (.50(.50)(.50)) \\
&= 1 - (\ .2500\ (.50)) \\
&= 1 - \ .125000 \\
\alpha_{EW} &= .875000.
\end{aligned}
$$

Table 2 illustrates these concepts more concretely. In the table are listed eight equally likely outcomes for sets of three coin flips. Of the eight sets of three flips, seven involve one or more Type I error, defined in this example as a heads. According to the Bonferroni inequality, 7/8 equals .875000, as expected.

As stated earlier, the above example is a literal fit for the Bonferroni inequality because the behavior of the coin on a given flip is <u>un</u>correlated with the behavior of the coin on any other flip. The exact α_{EW} can be determined using the Bonferroni inequality formula if the hypotheses or variables are perfectly uncorrelated. This formula is not necessary when the hypotheses are perfectly correlated because the α_{EW} and the α_{TW} equal each other.

However, in most studies hypotheses are neither perfectly uncorrelated nor perfectly correlated, and rather are partially correlated. For such studies, the <u>actual</u> experimentwise error rate will range somewhere between the computed experimentwise error rate (see above) and the testwise error rate, but may never really be known (Fish, 1988; Love, 1988; Morrorw & Frankiewicz, 1979).

Also, the α_{EW} inflation can be quite severe given the number of hypotheses tested and the level of correlation. Therefore, the power to reject can be low (Olejnik, Li, Supattathum, & Huberty, 1997). In other words, with multiple univariate follow-up tests at the original α_{TW} level (e.g., .05), the α_{EW} is inflated to statistical significance even if no statistical significance is found anywhere in the study. In order to compensate for this, researchers apply a "correction." This is called the "Bonferroni correction."

Bonferroni Correction

The Bonferroni correction compensates for the inflation by dividing the original α_{TW} by the number of k hypotheses in the study yielding a new α_{TW}*(Maxwell, 1992; Thompson, 1994):

$$\alpha_{TW}^{*} = \frac{\alpha_{TW}}{k}.$$

Each individual post hoc test then utilizes the α_{TW}* in order to maintain the α_{EW} at an appropriate level. Table 3 illustrates how the Bonferroni correction is utilized in order to maintain the α_{EW} at an appropriate level. However, this table also illustrates how the use of the Bonferroni correction has the potential for severe loss in power (Olejnik, Li, Supattathum, & Huberty, 1997).

Post Hoc Analysis

After using an ANOVA omnibus test to analyze overall differences in a multi-group study with more than two groups, many researchers use “post hoc” (also called “a posteriori,” “unplanned,” or “unfocused”) tests to determine which group means differ for each set of pairs or combinations of groups. All comparisons/contrasts only test whether exactly two means are equal. There are two kinds of comparisons: simple and complex. Although all contrasts test the equality of exactly two means, simple and complex contrasts differ as regards the permissible ways in which the two means are created. Put simply, "simple" contrasts compare the dependent variable means of two groups using the existing levels of a way, without any combinations of any levels. "Complex" contrasts, on the other hand, include all possible "simple" contrasts, but also include means computed by aggregating data across levels of the way.

For example, let's presume that a researcher did a one-way three-level ANOVA in which there were 10 people in each of the three groups of car owners: (a) Ford, (b) Nissan and (c) Rolls Royce. The dependent variable might be satisfaction with one's car. For this design three "simple" contrasts of mean levels of satisfaction are possible:

M_{FORD} (n = 10) = M_{NISSAN} (n = 10);

M_{FORD} (n = 10) = M_{ROLLS} (n = 10); and

M_{NISSAN} (n = 10) = M_{ROLLS} (n = 10).

The "complex" contrasts include these simple contrasts, plus the following three "uniquely complex" contrasts:

M_{FORD} (n = 10) = $M_{NISSAN\ or\ ROLLS}$ (n = 20);

M_{NISSAN} (n = 10) = $M_{FORD\ or\ ROLLS}$ (n = 20); and

M_{ROLLS} (n = 10) = $M_{FORD\ or\ NISSAN}$ (n = 20).

Table 4 illustrates these combinations for both three- and four-level one-way ANOVA problems. As Table 4 makes clear, as the number of levels gets larger, the number of simple contrasts gets larger, but the number of complex contrasts gets <u>exponentially</u> larger.

For each comparison, simple or complex, there are specific post hoc tests used. For simple comparisons the Tukey method, also called the HSD (honestly significant difference) test, is often used. For complex comparisons the Scheffé method is often used (Hinkle, Wiersman, & Jurs, 1998). Each of these method utilizes an analogue to the Bonferroni correction in order to maintain the α_{EW} at the a priori α level.

Tukey

The Tukey method is likely the most recommended and used procedure for controlling Type I error when making simple comparisons. The original Tukey method is based on Studentized range statistics, which takes into account the number of means being compared, adjusting for the total number of tests to make all simple comparisons. Later revisions of the Tukey method have demonstrated its robustness to violations of normality and homogeneity assumptions (Barnette, 1998). The Tukey method is also relatively insensitive to skewness. The Tukey method is not affected too much by many varied conditions. The exception to that is with the variability of the population means. Keselman (1976) found that the Tukey method is more powerful for the maximum variability of the population means. This is logical given that under this condition the magnitude of simple comparisons is largest. However, with larger sample sizes, the Tukey tends to lose relative power.

Scheffé

The Scheffé method is designed to analyze all possible comparisons (Sato, 1996). Therefore, the Scheffé method is used for complex or multiple comparisons. The Scheffé's infinite intersectional nature is its greatest strength and its greatest weakness. It is strong because it can analyze all possible comparisons. Klockars and Hancock (1998), however, assert that researchers are not always interested in many of the comparisons Scheffé makes. Because it is designed to test so many multiple comparisons, the Scheffé method is extremely conservative. The Scheffé methods suffers loss of power for some researchers because it is so conservative (Sato, 1996).

Multivariate Methods

Multivariate methods are designed for multiple outcome variables. As Huberty and Morris (1989) noted, multivariate methods ask, "Are there any overall effects present?" This questioning, or this philosophy, best honors the reality from which data are collected. That is, if data are collected from samples upon which there are many influences, or variables, then it is logical to use a statistical method that is designed to take those variables into account simultaneously (Thompson, 1994).

Because multivariate methods are designed for multiple outcome variables, multivariate methods require only one omnibus test to determine if any differences exist. This is in contrast to univariate methods, which require many tests, thus increasing the likelihood of making erroneous decisions. For this reason alone, multivariate methods should be used when multiple outcome variables are of concern.

Summary

Although many researchers are familiar with "testwise" alpha, "experimentwise" Type I error rates are also important, and must be considered in many research situations. Testing multiple hypotheses with a single sample of data can radically inflate the "experimentwise" Type I error rate.

The present paper has explained how this inflation can be avoided in various research situations. First, it was explained that ANOVA post hoc tests implicitly incorporate a hidden analog of the "Bonferroni correction" to avoid Type I error rate inflation. Second, it was noted that multivariate statistics are frequently employed by researchers to control "experimentwise" errors that would otherwise occur by conducting several ANOVA's or regression analyses with a single sample of data.

References

Barnette, J. J. (1998, November). The Tukey Honestly Significant Difference procedure and its control of the Type I error rate. Paper presented at the annual meeting of the Mid-South Educational Research Association, New Orleans. (ERIC Document Reproduction Services No. ED 427 043)

Fish, L. J. (1988). Why multivariate methods are usually vital. Measurement and Evaluation in Counseling and Development, 21, 130-137.

Hinkle, D. E., Wiersma, W., & Jurs, S. G. (1998). Applied Statistics for the Behavioral Sciences (4th ed.). New York: Houghton Mifflin Company.

Huberty, C. J, & Morris, J. D. (1989). Multivariate analysis versus multiple univariate analysis. Psychological Bulletin, 105, 302-308.

Kockars, A. J. & Hancock, G. R. (1998). A more powerful post hoc multiple comparison procedure in analysis of variance. Journal of Educational and Behavioral Statistics, 23, 279-289.

Love, G. (1988, November). Understanding experimentwise error probability. Paper presented at the annual meeting of the Mid-South Educational Research Association, Louisville. (ERIC Document Reproduction Service No. ED 304 451)

Maxwell, S. (1992). Recent developments in MANOVA applications. In B. Thompson (Ed.), Advances in social science methodology (Vol. 2, pp. 137-168). Greenwich, CT: JAI Press.

Morrow, J. R., & Frankiewicz, R. G. (1979). Strategies for the analysis of repeated and multiple measures designs. Research Quarterly, 50, 297-304.

Olejnik, S., Li, J., Supattathum, S., & Huberty, C. J. (1997). Multiple testing and statistical power with modified Bonferroni procedures. Journal of Educational and Behavioral Statistics, 22, 389-406.

Sato, T. (1996). Type I and Type II error in multiple comparisons. Journal of Psychology, 130, 293-302.

Thompson, B. (1994, April). Common methodology mistakes in dissertations, revisited. Paper presented at the annual meeting of the American Educational Research Association, New Orleans. (ERIC Document Reproduction Service No. ED 368 771)

Thompson, B. (1999, April). Common methodology mistakes in educational research, revisited, along with a primer on both effect sizes and the bootstrap. **Invited address** presented at the annual meeting of the American Educational Research Association, Montreal. (ERIC Document Reproduction Service No. ED 429 110)

Witte, R. S. (1980). Statistics. New York: Holt, Rinehart and Winston.

Table 1

Experimentwise Error Inflation Rates

α_{TW}	Tests	α_{EW}
1 - (1 - .01) **	1 =	
1 - (0.99) **	1 =	
1 - 0.99	=	0.01
1 - (1 - .001) **	10 =	0.009955
1 - (1 - .001) **	20 =	0.019811
1 - (1 - .001) **	30 =	0.029569
1 - (1 - .001) **	40 =	0.039230
1 - (1 - .001) **	50 =	0.048794
1 - (1 - .01) **	10 =	0.0561792
1 - (1 - .01) **	20 =	0.1820931
1 - (1 - .01) **	30 =	0.2602996
1 - (1 - .01) **	40 =	0.3310282
1 - (1 - .01) **	50 =	0.3949939
1 - (1 - .05) **	10 =	0.401263
1 - (1 - .05) **	20 =	0.641514
1 - (1 - .05) **	30 =	0.785361
1 - (1 - .05) **	40 =	0.871488
1 - (1 - .05) **	50 =	0.923055

Table 2

All Possible Families of Outcomes

for a Fair Coin Flipped Three Times

```
     Flip #
     1   2   3
 1.  T : T : T __
 2.  H : T : T   |  p of 1 or more H's (TW error analog)
 3.  T : H : T   |  in set of 3 Flips = 7/8 = 87.5%
 4.  T : T : H   |
 5.  H : H : T   |            or
 6.  H : T : H   |  where TW error analog = .50,
 7.  T : H : H   |  EW p = 1 - (1 - .5)^3
 8.  H : H : H __|     = 1 - (.5)^3
                       = 1 - .125 = .875
p of H on
each Flip      50% 50% 50%
```

Table 3

Experimentwise Error Rate Without and With The Application of the Bonferroni Correction

Number of Hypotheses	$1 - (1 - \alpha_{TW})^k$	=	α_{EW}
No Bonferroni Correction			
10	$1 - (1 - .05)^{10}$	=	.40126
50	$1 - (1 - .05)^{50}$	=	.92306
100	$1 - (1 - .05)^{100}$	=	.99408
Bonferroni Correction			
10	.05/10 =.00500		.04889
50	.05/50 =.00100		.04879
100	.05/100=.00050		.04878

Note. All original α_{TW} for equations in Table 3 are at the .05 level.

Table 4
List of Simple and Complex Contrasts
for One-way _k_=3 and _k_=4 ANOVA

Design   Contrasts

k=3 levels

Simple [3 (3 - 1)] / 2 = 6 / 2 = **3**
1 vs 2
1 vs 3
2 vs 3

Complex 3 + 3 = **6**

Simple
1 vs 2
1 vs 3
2 vs 3

Uniquely complex
1 vs 2,3
2 vs 1,3
3 vs 1,2

k=4 levels

Simple [4 (4 -1)] / 2 = 12 / 2 = **6**
1 vs 2
1 vs 3
1 vs 4
2 vs 3
2 vs 4
3 vs 4

Complex 6 + 15 = **21**

Simple
1 vs 2
1 vs 3
1 vs 4
2 vs 3
2 vs 4
3 vs 4

Uniquely complex
1, 2 vs 3
1, 2 vs 4
1, 3 vs 4
2, 1 vs 3
2, 1 vs 4

2, 3 vs 4
3, 1 vs 3
3, 1 vs 4
3, 2 vs 4
4, 1 vs 2
4, 1 vs 3
4, 2 vs 3
1, 2 vs 3, 4
1, 3 vs 2, 4
1, 4 vs 2, 3