DIGITALCOMMONS
— @WAYNESTATE —

Journal of Modern Applied Statistical Methods

Volume 5 | Issue 2 Article 23

11-1-2005

Corrections for Type I Error in Social Science Research: A Disconnect between Theory and Practice

Kenneth Lachlan
Boston College, lachlan@bc.edu

Patric R. Spence
Western Kentucky University

Recommended Citation

Lachlan, Kenneth and Spence, Patric R. (2005) "Corrections for Type I Error in Social Science Research: A Disconnect between Theory and Practice," *Journal of Modern Applied Statistical Methods*: Vol. 5 : Iss. 2 , Article 23.
DOI: 10.22237/jmasm/1162354920
Available at: http://digitalcommons.wayne.edu/jmasm/vol5/iss2/23

Journal of Modern Applied Statistical Methods
November, 2006, Vol. 5, No. 2, 490-494

1538 – 9472/06/$95.00

Corrections for Type I Error in Social Science Research: A Disconnect between Theory and Practice

Kenneth Lachlan
Department of Communication
Boston College

Patric R. Spence
Department of Communication
Western Kentucky University

Type I errors are a common problem in factorial ANOVA and ANOVA based analyses. Despite decades of literature offering solutions to the Type I error problems associated with multiple significance tests, simple solutions such as Bonferroni corrections have been largely ignored by social scientists. To examine this discontinuity between theory and practice, a content analysis was performed on 5 flagship social science journals. Results indicate that corrections for Type I error are seldom utilized, even in designs so complicated as to almost guarantee erroneous rejection of null hypotheses.

Key words: Type I error, false positive, Bonferroni, Type II error

Introduction

Despite the breadth of literature in statistical and methodological research detailing the problems associated with Type I error and multiple F tests in factorial ANOVA (Fletcher, et al. 1989, Keppel, 1991, Cohen, 1994, Agresti & Finlay, 1997, Mulaik, Raju, & Harshman, 1997; Smith et al., 2002, Padilla & Algina, 2004), a cursory examination of social science literature suggests that these warnings have been largely ignored. This article briefly reviews some of the literature concerning Type I error rates, then offers an ad hoc content analysis of several leading journals in different social science disciplines. The results of this content analysis suggest that there is a serious split between statistical literature warning researchers about the Type I error problems associated with multiple F tests in factorial ANOVA, and the actual practice of statistical inference in social scientific research.

Type I errors refer to instances in which a null hypothesis is erroneously rejected. Type I error may be the result of several factors (such as a high alpha level or the violation of statistical assumptions), but the most common source appears to be the number of significance tests that are calculated (Steinfatt, 1979). Although it is well documented that multiple tests along different levels of a single factor will produce Type I errors, less documented is the fact that multiple F tests alone will increase the probability of Type I error (Fletcher, et al., 1989). When testing at the commonly accepted criterion of $p < .05$, one out of every twenty tests will produce an error of Type I (assuming the null hypothesis is always true). Calculations can be performed to compute the expected

Ken Lachlan is Assistant Professor in the Communication Department at Boston College. He earned his doctoral degree from Michigan State University in 2003 email: lachlan@bc.edu. Patric Spence is Assistant Professor in the Department of Communication at Western Kentucky University. He earned his doctoral degree from Wayne State University in 2005.

probability of Type I error through the equation $1 - (1 - \alpha)^c$ where c represents the number of independent comparisons. (Keppel, 1991, Steinfatt, 1979, Smith et al., 2002).

The most commonly used correction for Type I error is a simple reduction of alpha, usually through Bonferroni corrections. These corrections divide the alpha level by the number of tests being performed, then set each test accordingly (Agresti & Finlay, 1997, Cohen & Cohen, 1983, Keppel, 1991).

Fletcher et al. (1989) performed a series of Monte Carlo simulations demonstrating a substantial increase in the number of Type I errors corresponding with the number of factors in a given model. With regard to Bonferroni corrections, Fletcher et al. (1989) reported that Type I error rates dropped from 32 percent to 11 percent through the use of these corrections, using a three-factor ANOVA model in which the null was assumed to be true.

Smith et al. (2002) attempted to extend the work of Fletcher and colleagues by conducting a series of similar Monte Carlo simulations using three and four factor models in which the null is sometimes assumed true and sometimes assumed false. They reported that the addition of main effects into multi-factor models, the use of larger samples, and Bonferroni corrections substantially reduce Type I error rates, to levels as low as 2% across 500 trial models. They caution, however, that Bonferroni corrections may in fact be too conservative and in turn inhibit the detection of true effects, increasing errors of Type II.

Concern over the hypersensitivity of Bonferroni corrections is nothing new. Simes (1986), Hochberg (1988), and Hommel (1988) offer more mathematically sophisticated means of adjusting alpha levels based on sequential adjustments relative to the number of tests that have been performed, rather than the total number of tests performed on a given model. Keppel (1991) offered a modified Bonferroni adjustment that is based on the number of groups used in the model, as opposed to the total number of tests. Monte Carlo simulations of these techniques demonstrate their effectiveness, and they have been lauded for their ability to effectively reduce Type I error without excessive Type II risk (e.g. McDonald, Seifert, Lorenzet, Givens, & Jaccard, 2002).

As outlined, a substantial body of research has been devoted to identifying and correcting for Type I errors in social science research. Although scholars in applied statistics have debated whether to use the original Bonferroni formula or some type of adjusted formula, the fact remains that correction for Type I errors across multiple tests in multi-factor ANOVA has been identified as a necessary and important component of factorial inference. Without consideration of Type I error, statistical conclusion validity (see Cook & Campbell, 1979) is called into question, with grave implications for the usefulness and validity of findings that are based solely on estimations of the likelihood that they are false (Nickerson, 2000).

However, it is likely that the reader can think of dozens of articles he or she has read recently which have used multi-factor ANOVA procedures and performed numerous F tests, with no regard for Type I errors or the necessary adjustments. Indeed, Smith et al. (2002) in a review of Communication research report that about a quarter of the articles examined featured ANOVA designs of 3 or more factors, with almost none adjusted for the error rates produced by multiple F tests.

The goal of the current analysis to examine a few major journals in the social sciences in order to obtain an estimate of the frequency with which Type I corrections-

Bonferroni or otherwise- are considered and implemented in contemporary research. To do so, a content analysis was performed by the authors examining quantitative research articles in each of the following journals during the 2004 calendar year: Journal of Personality and Social Psychology, Personality and Individual Differences, Human Communication Research, Educational and Psychological Measurement, and the American Journal of Public Health.

Methodology

An initial examination of all articles appearing in these journals during the 2004 calendar year was conducted in order to identify articles using some kind of ANOVA or related analysis. A total of 6 articles were identified among 423 articles appearing in AJPH (1.42%); 36 out of 58 were found for JPSP (62.1%), 4 of 61 (6.6%) for EPM, 10 of 22 (45.5%) for HCR, and 96 of 296 (32.4%) for PID.

Two coders were then given the task of coding several content features of each article. Specifically, they were asked to identify whether or not the article reported ANOVA, ANCOVA, MANOVA, or MANCOVA procedures, the total number of analyses, the number of F tests reported, an estimate of the largest single cell size across all analyses, and whether or not Bonferroni or forms of Type I error correction were employed. Intercoder reliability was calculated using Scott's Pi for categorical variables and Kronbach's alpha for continuous variables; reliability checks on 10% of the sample produced coefficients of at least .87 for all variables.

It should be noted that for definitional purposes coding was completed solely for the number of F tests reported, not an estimated number of total possible F tests. This decision was made for two reasons: first, to produce a conservative estimate of the number of F tests that were run in each study; and second, because there were numerous instances in which the statistical reporting was so ambiguous that it was impossible to estimate the total number of tests that could have been run.

Results

Results indicate that Bonferroni and other corrections for Type I error are generally absent in these journals, as only 15.8% of the identified articles reported such a correction. More specifically, only 2.8% of the identified JPSP articles reported these corrections, along with 10% of HCR, 16.7% of AJPH, 20.8% of PID, and 25% of the EPM articles. It is perhaps not surprising that Educational and Psychological Measurement, a journal in which psychometric pieces are quite common, would have the highest incidence of reports in which Type I error corrections were performed. But the general tenor of these findings is that Bonferroni and similar procedures are under reported in social science literature.

These results would not be particularly alarming if the studies featured in these journals performed a small number of F-tests with *p* set at .05. However, there were numerous instances in which this was not the case. Across the entire sample, the average study contained 5.86 ANOVA or ANOVA related analyses, and the average number of reported F-tests was found to be 14.51. Given that a *p* value criterion of .05 should produce one false positive out of every 20 tests by chance alone, simple frequency distributions were used to determine the number of articles reporting 20 or more tests; in total, 34 of 152 (22.4%) or the articles reported enough F-tests without corrections that at least one Type I error could be expected. A few were particularly notable, including one study that featured an incredible 111 F-tests.

An argument could be made that ANOVA procedures in the social sciences experience a higher incidence of Type II errors due to small sample sizes and resultant lack of statistical power, thus justifying numerous F-tests without oversensitive corrective measures. Although the authors would argue that this logic leads to errors of both Type I and Type II and actually leads to even less statistical conclusion validity, the concerns associated with underpowered analyses can be seen. However, this analysis suggests that at least in these journals, Type II errors are likely less of a concern.

Across the entire sample, the median score for maximum cell size was found to be 28.5. While there are obviously varying scenarios in which this may present an adequately or inadequately powered analysis, it at least is above the minimum criteria set for adequate cell size in ANOVA analysis It also needs to be noted that the mean score for maximum cell size (104.6) was intensely skewed by a handful of epidemiological studies with samples of over one thousand. For this reason, the median score is reported, which the authors believe to be a better indicator of central tendency.

Another logical question concerns differences between the studies utilizing error corrections and those that do not in terms of the number of reported F-tests. It could be argued that if the studies running dozens of F-tests are the ones controlling for Type I error, then there is no cause for consternation. This is, however, not the case. *T* tests were used to examine differences in these scores. For reported F-tests, significant differences were not detected between those studies using Type I corrections (*M* = 11.5, *SD* = 8.81) and those that did not (*M* =15.08, *SD* = 18.38), *t*(150) = 3.91, n.s.

Conclusion

The results reported above, while confined to only a few journals in one calendar year, suggest a disconnect between the statistical analysis and reporting procedures commonly advocated in the statistics literature and the actual practices of social scientists. Across this sample of flagship journals in Public Health, Communication, Psychology, and Education, techniques for reducing Type I error in factorial ANOVA that have been advocated, debated, and refined in the statistics literature for decades are going largely unused. Although this may be less of a concern in the Education and Public Health literatures, empirical Psychology and Communication is largely dependent on ANOVA analyses, especially when utilizing experimental designs.

Further, it should be noted that the statistical reporting under scrutiny was often composed of several sets of multi-factor analyses, leading to situations in which dozens of F-tests were reported. In several studies it could be extrapolated that the null hypothesis may have been erroneously rejected upwards of four times simply due to chance. The most plausible solution is to follow years of advice from the statistics literature and correct alpha with Bonferroni or other adjustments for Type I error, allowing the researcher to differentiate between statistically valid findings and falsely rejected null hypotheses. Although some have suggested that Bonferroni corrections and other adjustments may be overly sensitive and in fact lead to an increase in Type II error (Smith et al., 2002), this increase in Type II error can be avoided through the use of larger samples.

Although obtaining large samples for experimental studies can often be difficult, costly, and time-consuming, it is the opinion of the researchers that the best possible solution is to use enough subjects to provide for adequate statistical power (see Keppel, 1991) and perform corrections for Type I errors. In instances in which this is not possible it is recommended that social scientists report their effect sizes and establish an a priori criterion for findings that will be considered relevant based on effect size (see Cohen, 1977 for suggested effect size criteria).While not a formal part of this study's coding scheme, it should be noted that the coders observed many articles in which multiple F-tests were reported with no regard for effect size. Reporting only those findings that are statistically significant at .05 and that meet an established criterion for the amount of variance accounted for may be one more solution to the prevalence of Type I error.

References

Agresti, A., & Finlay, B. (1997). *Statistical methods for the social sciences* (3rd Edition). Prentice Hall.

Cohen, J. (1994). The earth is round (p<.05). *American Psychologist, 49,* 997-1003.

Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences.* Hillsdale, N.J.: Erlbaum.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences*. New York: Academic Press.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-Experimentation*. Boston: Houghton Mifflin.

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

Hommel, G. (1988). A stagewise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika, 75,* 383-386.

Keppel, G. (1991). *Design and analysis: A researcher's handbook*. Upper Saddle River, N.J.: Prentice Hall.

McDonald, R. A., Seifert, C. F., Lorenzet, S. J., Givens, S., & Jaccard, J. (2004). The effectiveness of methods for analyzing multivariate factorial data. *Organizational Research Methods, 5,* 255-274.

Mulaik, S. A., Raju, N. S., & Harshman, R. A. (1997). There is a time and place for significance testing. In L.L. Harlow, S.A. Mulaik, & J.H. Steiger (Eds.) *What if there were no significance tests?* Mahwah, N.J.: Earlbaum.

Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods, 5*, 241-301.

Padilla, M. A., & Algina, J. (2004). Type I error rates for a one-factor within-subjects design with missing values. *Journal of Modern Applied Statistical Methods, 3,* 406-416.

Simes, R. J. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika, 73,* 751-754.

Smith, R. A., Levine, T. R., Lachlan, K. A., & Fediuk, T. A. (2002). The high cost of complexity in experimental design and data analysis: Type I and Type II error rates in multiway ANOVA. *Human Communication Research, 28*, 515-530.

Fletcher, H. J., Daw, H., & Young, J. (1989). Controlling multiple F test errors with an overall F test. *Journal of Applied Behavioral Science, 25*, 101-108.

Steinfatt, T. M. (1979). The alpha percentage and experimentwise error rates in communication research. *Human Communication Research, 5,* 366-374.