ISSN 1750-9823 (print)
International Journal of Sports Science and Engineering
Vol. 07 (2013) No. 01, pp. 017-020

Alpha Level Adjustments for Multiple Dependent Variable Analyses and Their Applicability – A Review

Sinclair, J.K [1], Taylor, P.J [2,] and Hobbs, S.J [1].

[1] School of Sport Tourism and Outdoors, University of Central Lancashire
[2] School of Psychology, University of Central Lancashire

(*Received April 10, 2012, accepted September 9, 2012*)

Abstract. Researchers in sport science typically find themselves simultaneously examining many questions and hypothesis through multiple dependant variable analyses. When more than one independent test is used, the likelihood of finding significant effects due to chance increases linearly based on the number of analyses being conducted. In recent years however a division has emerged between those performing comparative clinical or semi-clinical analyses with aetiological components and so called conventionalists who have continued to employ traditional methods such as Bonferroni corrections to control for type I error. The problem with alpha level adjustments is that whilst they do reduce the likelihood of making a type I error, the probability of making a type II error correspondingly increases. This review advocates a strategy of not making adjustments for multiple analyses as it leads to less errors of interpretation. Researchers should not be punished by missing potentially meaningful findings for their willingness to explore additional information. Sport science by its very nature comprises a multitude of hypotheses and comparisons, and this simple fact leads to the conclusion that adjustments for multiple comparisons are not necessary.

Keywords: Multiple comparisons, statistical significance, type I error, type II error

1. Introduction

The goal of comparative inferential statistical analyses in sports science is to make inferential decisions regarding the effect of two or more independent variables on an outcome measure in relation to the null hypothesis (Thomas and Nelson, 2005). Researchers typically find themselves simultaneously examining many questions and hypothesis through multiple dependant variables analyses (Hsu, 1996 and Ludbrook, 1991). The key issue surrounding multiple comparisons when using an alpha level of 0.05 is type I error, whereby a researcher may erroneously reject the null hypothesis due to the number of analyses being made (Benjamini and Hochberg, 1995 and Shaffer, 1995).

In recent years a division has emerged between those performing comparative clinical and semi-clinical analyses with aetiological components, and so called conventionalists who have continued to employ traditional methods such as Bonferroni corrections to control for type I error (Rothman, 1990 and Feise, 2002). Both sides have presented legitimate arguments to support their approaches, but a consensus has yet to be reached. The primary aim of this review is to evaluate the need for alpha level adjustments when multiple comparisons are made.

2. Multiple comparisons and type I error – the traditional approach

Standard academic practice for statistical analyses is to accept an alpha level of 0.05 in order to distinguish statistical significance from non-significance. By definition this approach when conducting twenty analyses will result in one variable that will appear to be significant when in reality it is co-incidental (Williams, 1971). The occurrence of rejecting the null hypothesis when it is in fact true is referred to as a type I error. When more than one independent test is used, the likelihood of finding one significant effect due to chance increases linearly based on the number of tests that are conducted (Williams, 1971).

Therefore alpha level adjustments are based around the following premise: if the null hypothesis is indeed true then significant observations may still be observed. To accommodate and control for this, the alpha level for each analysis is adjusted to ensure that the overall likelihood of obtaining a significant effect is still at the α=0.05 level. This importantly allows multiple analyses to be performed with a minimal risk of

type I error. Conventionalists who advocate adjustments for multiple comparisons argue that control over type I error or false positives are imperative to avoid spurious associations (Perneger, 1998).

3. General procedures for multiple test adjustments

Classicists believe that if multiple measures are tested in a given study, the alpha level should be adjusted in order to reduce the chance of observing spurious statistical significance (Tukey, 1977, Bland and Atman 1995). This view is based on the theory that if you test long enough, you will inevitably find something statistically significant (false-positives) due to random variability, even if no real effects exist (Greenhalgh, 1997 and Ludbook, 1998). This has been referred to as the multiple testing problem or the problem of multiplicity (Ahlbom, 1993). A variety of methods have been developed, but no gold standard method exists (Sidak, 1967, Williams, 1971 and Holm, 1979).

3.1. Bonferroni

One of the fundamental and historically utilized adjustments for type I error is the Bonferroni correction (Perneger, 1998). The Bonferroni correction adjusts the alpha level at which a statistical test considered to be significant based on the total number of analyses being conducted. Specifically, the utilized alpha level is quantified as being the original alpha level of α=0.05 divided by the number of comparisons being made. Implicitly, the Bonferroni adjustment assumes that these test statistics are independent. For example when conducting four analyses an overall desired alpha level of 0.05 would translate into individual tests each using an p-value threshold of 0.05/4 = 0.0125. The Bonferroni adjustment procedure has the advantage of being simple and valid even when the analyses being conducted are dependent. Although Bonferroni is the conventional method of adjusting the alpha level, it is frequently considered to be overly conservative.

3.2. Holm-Bonferroni

An equivalently more powerful statistical procedure (i.e. more stringent irrespective of the values of the unobservable parameters) is the Holm–Bonferroni technique (Holm, 1979). The Holm-Bonferroni technique is a successively rejective adaptation of the simpler Bonferroni adjustment for multiple analyses, and strongly controls the alpha level. The Holm-Bonferroni adjustment ranks all of the observed p-values in order from smallest to largest, if the first p-value is greater than or equal to the alpha level/ the number of comparisons being made then the procedure is halted and the null hypothesis is accepted. If the first p-value is found to be significant (i.e. less than p-value/ the number of comparisons) the second p-value is compared to the alpha level / number of comparisons-1. This process continues until one of the variables is found to be non-significant then the analysis ceases.

3.3. Hochberg

The Hochberg procedure is very similar to the Holm-Bonferroni technique (Hochberg, 1988). The only difference is that this technique is regarded as a step-up, rather than step-down, procedure as it ranks the observed p-values from high to low. The analysis then examines the observed p-values from the highest to the lowest, and discontinues as soon as the p-value is less than is the adjusted alpha, and from here onwards represents significant p-values.

3.4. Sidak

The Šidák correction is conducted by assuming that the conducted analyses tests are independent (Sidak 1967). Since all of the variables are considered to be independent, the adjusted alpha level is equal to 1-(1-unadjusted alpha level) multiplied by the number of comparisons. The Šidák correction gives a stronger bound than the Bonferroni correction but can be limited by the condition of independence and is less stringent in its control over type I error. Because the Šidák correction requires calculating fractional powers, it is complicated to perform and the simpler Bonferroni correction is often preferred.

3.5. Dunnett

Dunnett's test is specifically designed to allow variables means to be contrasted against a single reference mean (Dunnett, 1955 and 1964). It is commonly used after the homogeneity assumption has been violated. Its aim is to identify variables whose means are significantly different from the reference. It examines the null hypothesis in that none of the variable means is significantly different from the reference mean.

4. Abandoning the type I error paradigm

Whilst the justification for techniques such as Bonferroni corrections appears to be reasonable, two key empirical issues exist. Alpha level adjustment procedures such as Bonferroni serve to examine a so called universal null hypothesis against the alternative hypothesis (O'Keefe, 2003 and Savitz and Olshan 1995). As such the rejection of the universal null hypothesis as opposed to the alternative is simply a statement that one or more of the variables that compromise the universal null hypothesis is rejected, but without the ability to define which one (Cox and Wong, 2004 and Wacholder et al., 2004). It could be argued that a researcher should be most interested in examining individual hypotheses, and that examining the so called composite hypothesis is rarely of practical or scientific concern. A second more understated problem for researchers and statisticians is that the likelihood of a falsely rejected null hypothesis cannot be localized to an unambiguous set of analyses (Everitt, 2000). Simply stated a researcher is able to subjectively select the analyses over which an alpha level adjustment is applied, and this subjective choice can produce unreliable conclusions.

A further objection that researchers typically make to alpha level adjustments is that whilst the likelihood of making a type I error is reduced; the probability of making a type II error correspondingly increases (Rothman, 1990, Perneger, 1998 and Thomas et al., 1985). By changing the alpha level required to reject the null hypothesis (or equivalently widening the uncertainty intervals) the quantity of rejected null hypotheses will decrease (Halperin et al., 1988 and Einot, and Gabriel, 1975). Although this will serve to reduce the number of false rejections, it will also serve to increase the number of instances in which the null hypothesis is not rejected when in fact it is false. As such, the conventionally advocated alpha level corrections can severely reduce the power to detect an important effect. By reducing the likelihood of type I errors through alpha level adjustments, you increase the incidence of type II error. Type II errors can be no less serious than type I errors particularly in sport science as it may result in a valid result being discarded. Thus, it is recommended that the consequences of type II errors be considered more extensively by researchers.

5. Discussion

The proposed approach advocated by this article, has two fundamental differences from the classical perspective. Firstly, it is recommended that Type 1 error be de-emphasized to some extent because it is not possible for the null hypothesis to be strictly accurate. Secondly, it is proposed in sports science that type II errors pose a more significant threat to the efficacy of exploratory analyses; and that researchers should not be punished for presenting a more complete picture of their study through the inclusion of more variables. Conventionalists who subscribe to the theory of adjustments for multiple analyses face the problem described by Rothman, (1990) as the penalty for peeking. If this premise is allowed, many logical inconsistencies may arise.

The paradox of paying a penalty for having more information is a concept that has commonly been accepted (Rothman, 1990 and Benjamini and Hochberg, 1995). The paradox arises only if researchers are willing to assume the truth of the universal null hypothesis; however, the premise of a universal null hypothesis is one that empirical science constantly refutes (Altman, 1991 and Armitage and Perry, 1994) as it lacks any apparent heuristic value. Therefore to pay a penalty for making additional observations should be considered unacceptable to any scientist. This review advocates a strategy of not making adjustments for multiple analyses as it will lead to less errors of interpretation. Researchers should not be punished by missing potentially meaningful findings for their willingness to explore additional information. Sport science by its very nature comprises a multitude of hypotheses and comparisons, and this simple fact leads to the conclusion that adjustments for multiple comparisons are not necessary.

6. References

[1] Altman D.G. *Practical Statistics for Medical Research*. New York: Chapman & Hall. 1991, pp. 210-212.

[2] Armitage P. and Berry G. *Statistical Methods in Medical Research*(3rd ed). Cambridge, MA: Blackwell Scientific Publications. 1994, pp. 224-228.

[3] Ahlbom A. *Biostatistics for Epidemiologists*. Boca Raton: Lewis Publishers. 1993, pp. 52-53.

[4] Benjamini and Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society.*1995, **57**: 289-300.

[5] Bland J.M, Altman D.G. Multiple significance tests: the Bonferroni method. *British Medical Journal.* 1995, **310**: 170.

[6] Cox D.R, Wong M.Y. A Simple Procedure for the Selection of Significant Effects. *Journal of the Royal Statistical*

Society. 2004, **66**: 395-400.

[7] Douglas C.E. Multiple comparisons: philosophies and illustrations. *American Journal of Physiology – Regulatory, Integrative and Comparative Physiology*. 2000, **279**: 1-8.

[8] Dunnett, C.W. A multiple comparison procedure for comparing several treatments with a control. *Journal of the American Statistical Association.* 1955, **50**: 1096-1121

[9] Dunnett, C. W. New tables for multiple comparisons with a control. *Biometrics.* 1964, **20**: 482-491

[10] Einot, I. and Gabriel, K.R. A study of the powers of several methods of multiple comparisons. *Journal of the American Statistical Association.* 1975, **70**: 574-583.

[11] Feise R.J. Do multiple outcome measures require p-value adjustment? *BMC Medical Research Methodology.* 2002, **2**: 2-8

[12] Greenhalgh T. Statistics for the non-statistician. Different types of data need different statistical tests. *British Medical Journal*. 1997, **315**: 364-366.

[13] Halperin, M., Lan, K.K. and Hamdy, M.I. Some implications of an alternative definition of the multiple comparison problem. *Biometrika.* 1988, **75**: 773-778.

[14] Hochberg Y. A sharper Bonferonni procedure for multiple tests of significance. *Biometrika.* 1988, **75**: 800–803.

[15] Hsu JC. *Multiple Comparisons: Theory and Methods*. New York: Chapman & Hall, 1991.

[16] Ludbrook J. On making multiple comparisons in clinical and experimental pharmacology and physiology. *Clinical Experimental Pharmacology and Physiology*. 1996, **18**: 379-392.

[17] Ludbrook J. Multiple comparison procedures updated. *Clinical Experimental Pharmacology and Physiology.* 1998, **25**: 1032-1037.

[18] Holm S. A simple sequentially rejective multiple test procedure. *Scandinavian Journal Statistics.* 1979, **6**: 65-70.

[19] O'Keefe D.J. Should family-wise alpha be adjusted? *Human Commutative Research*. 2003, **29**: 431-447.

[20] Perneger T.V. What's wrong with Bonferroni adjustments. *British Medical Journal.* 1998, **316**: 1236-1238.

[21] Rothman K.J. No adjustments are needed for multiple comparisons. *Epidemiology.* 1990, **1**: 43-46.

[22] Savitz D.A, Olshan A.F. Multiple Comparisons and Related Issues in the Interpretation of Epidemiologic Data. *American Journal of Epidemiology*. 1995, **142**: 904-908.

[23] Shaffer, J.P. Multiple Hypothesis Testing. *Annual Review of Psychology.*1995, 46: 561-584.

[24] Sidak Z. Rectangular confidence regions for the means of multivariate normal distribution. *Journal of American Statistical Association.* 1967, 62: 626-633.

[25] Tukey J.W. Some thoughts on clinical trials, especially problems of multiplicity. *Science.* 1977, **198**: 679-684.

[26] Thomas, J.R and Nelson J.K *Research Methods in Physical Activity 5th Ed*. Human Kinetics, 2005. .

[27] Thomas D.C, Siemiatycki J., Dewar R., Robins J., Goldberg M., Armstrong B.G. The problem of multiple inference in studies designed to generate hypotheses. *American Journal of Epidemiology.* 1985, **122**: 1080-1095.

[28] Wacholder S., Chanock S. and Garcia-Closas M. Assessing the probability that a positive report is false: An approach for molecular epidemiology studies. *Journal of the National Cancer Institute*. 2004, **96**:434-442.

[29] Williams D.A. A test for differences between treatment means when several dose levels are compared with a zero dose control. *Biometrics*. 1971, **27**: 103-117.