ROBUST MULTIPLE PAIRWISE COMPARISON PROCEDURE FOR ADAPTIVE TRIMMED MEAN VIA P-METHOD

LOW JOON KHIM

MASTER OF STATISTICS
UNIVERSITI UTARA MALAYSIA
2016

Permission to Use

In presenting this thesis in fulfilment of the requirements for a postgraduate degree from Universiti Utara Malaysia, I agree that the Universiti Library may make it freely available for inspection. I further agree that permission for the copying of this thesis in any manner, in whole or in part, for scholarly purpose may be granted by my supervisor(s) or, in their absence, by the Dean of Awang Had Salleh Graduate School of Arts and Sciences. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to Universiti Utara Malaysia for any scholarly use which may be made of any material from my thesis.

Requests for permission to copy or to make other use of materials in this thesis, in whole or in part, should be addressed to:

Dean of Awang Had Salleh Graduate School of Arts and Sciences
UUM College of Arts and Sciences
Universiti Utara Malaysia
06010 UUM Sintok

Abstrak

Ujian omnibus teguh yang boleh didapati secara meluas biasanya digunakan sebagai alternatif kepada Analisis Varians (ANOVA) klasik apabila andaian tidak dipenuhi. Seperti ANOVA, setiap ujian omnibus memerlukan prosedur *post hoc* (perbandingan pasangan berganda) apabila ujian didapati signifikan. Walau bagaimanapun, kajian terhadap prosedur *post hoc* untuk ujian omnibus teguh yang sedia ada kurang diberi perhatian. Kebanyakan ujian omnibus teguh dibiarkan tanpa prosedur *post hoc* dan ujian sebegini dianggap tidak lengkap. Dalam kajian ini, kami telah mengambil inisiatif untuk membangunkan prosedur *post hoc* yang dikenali sebagai Kaedah-*P* untuk *HQ* dan HQ_1, iaitu dua penganggar teguh priori yang digunakan dalam menguji kesamaan kumpulan. Selain daripada dua penganggar teguh tersebut, kajian ini juga mengkaji keberkesanan min klasik menggunakan Kaedah-*P*. Kaedah-*P* adalah kaedah yang berasaskan bootstrap. Masing-masing ditandakan sebagai *P*-*HQ*, P-HQ_1 dan *P*-Min, program komputer untuk prosedur tersebut telah dibangunkan dan keberkesanannya dalam mengawal ralat Jenis I (keteguhan) telah dinilai. Satu kajian simulasi telah dijalankan untuk mengkaji kekuatan dan kelemahan prosedur. Bagi tujuan tersebut, lima pembolehubah telah dimanipulasikan untuk mewujudkan pelbagai keadaan yang sering berlaku dalam kehidupan sebenar. Pembolehubah tersebut adalah bentuk taburan, bilangan kumpulan, saiz sampel, tahap kepelbagaian varians dan pasangan saiz sampel dan varians. Sebanyak 2000 set data telah disimulasi menggunakan pakej SAS/IML Versi 9.2. Kriteria teguh liberal Bradley telah digunakan sebagai penanda aras keteguhan setiap prosedur. Akhir sekali, kaedah yang dicadangkan (*P*-*HQ* dan P-HQ_1) dan *P*-Min dibandingkan dengan kaedah *LSD-Bonferroni Correction* yang sedia ada. Hasil kajian mendapati *P*-*HQ* dan P-HQ_1 berkesan mengawal ralat Jenis I dan dengan itu boleh digunakan sebagai prosedur *post hoc* untuk ujian omnibus yang didapati signifikan membabitkan penganggar *HQ* dan HQ_1. Di samping itu, kajian ini juga mendapati bahawa *P*-Min adalah teguh walaupun di bawah pelanggaran yang teruk. Kajian ini secara keseluruhannya berjaya menghasilkan ujian post hoc yang boleh percaya untuk penganggar *HQ* dan HQ_1.

Kata Kunci: Ujian *post hoc*, *P*-*HQ*, P-HQ_1, *P*-Min

Abstract

Robust omnibus tests which are widely available are commonly used as alternatives to the classical Analysis of Variance (ANOVA) when the assumptions are violated. Like ANOVA, each of these omnibus tests needs a post hoc (pairwise multiple comparison) procedure when the test turns out to be significant. However, works on post hoc procedures for the existing robust omnibus tests are not given much attention. Most of the robust omnibus tests are left without the post hoc procedures and the tests are deemed incomplete. In this study, we have taken the initiative to develop the post hoc test known as *P*-Method for *HQ* and HQ_1, the two robust estimators priori used in testing the equality of groups. Apart from the two robust estimators, this study also looked into the effectiveness of the classical mean using *P*-Method. *P*-Method is a bootstrap based method. Respectively denoted as *P-HQ*, $P\text{-}HQ_1$ and *P*-Mean, computer programs for the procedures were developed and their effectiveness in controlling Type I error (robustness) was evaluated. A simulation study was conducted to investigate on the strength and weakness of the procedures. For such, five variables were manipulated to create various conditions that often occur in real life. These variables are the shape of the distributions, number of groups, sample sizes, degree of variance heterogeneity and pairing of sample sizes and variances. A total of 2000 datasets were simulated using SAS/IML Version 9.2. Bradley's liberal criterion of robustness was adopted to benchmark each procedure. Finally, the proposed methods (*P-HQ* and $P\text{-}HQ_1$) and *P*-Mean were compared with the existing LSD-Bonferroni correction. The finding revealed that *P-HQ* and $P\text{-}HQ_1$ could effectively control Type I error and thus could be used as the post hoc procedure for significant omnibus test using *HQ* and HQ_1 estimators. In addition, this study also observed that *P*-Mean is robust even under severe violation of assumptions. In general, this study managed to develop a reliable post hoc test for *HQ* dan HQ_1 estimators.

Keywords: Post hoc test, *P-HQ*, $P\text{-}HQ_1$, *P*-Mean

Acknowledgement

Foremost, I would like to express the deepest appreciation to my first supervisor Assoc. Prof. Dr. Sharipah Soaad Bt. Syed Yahaya for the continuous guidance, caring and patience in my master study and research. I would also like to thank her for her continuous help and support in all stages of this thesis. Next, I would like to express my gratitude to my second supervisor Dr. Suhaida Bt. Abdullah for the useful comments and engagement in my master learning process. Without their supervision and constant help, this dissertation would not have been possible.

I would also like to thank my parents Low Kim Hong and Low Yok Em for always supporting and encouraging me with their best wishes and effort. The appreciation also goes to my husband Ch'ng Chee Keong for his emotional support, love and motivation during my study.

Special thanks to the various people in School of Quantitative Sciences, Universiti Utara Malaysia as they provided me a very useful and helpful assistance.

Thanks in million again to all for providing me a loving environment to complete my master study.

Table of Contents

List of Tables

List of Figures

List of Appendices

CHAPTER ONE
INTRODUCTION

1.1 Background

When group means are compared, then the null hypothesis of equality (or homogeneity) is rejected, at this point there is no equality among them, but we have no idea about the form of the inequality. Usually, we undertake an analysis thoroughly of the nature of the difference. For example which group mean(s) differ from the others or does mean of group 1 differ from that of group 2? Thus, multiple pairwise comparison procedure (MCP) is needed to answer these queries. Cause of rejection of the null hypothesis will be investigated by the MCPs. There are several powerful MCPs that we can use after observing experimental results. Since each MCP has its strengths and weaknesses, it is advisable to make comparison among the MCPs and choose the MCP which can control Type I error, as well as to maximize power. The most widely used MCPs and can be found in major statistical packages are procedures such as Least Significant Difference (LSD), Scheffé, Tukey, and Bonferroni. However, the procedures are adversely affected by nonnormality, particularly when variances are heterogenous and group sizes are unequal (Keselman, Cribbie & Wilcox, 2002). Under these conditions, the rate of Type I error will increase, and cause spurious rejections of null hypothesis, and power is reduced, resulting in the test effects going undetected. Actual Type I error can exceed or below the nominal level when the sample sizes are twenty or smaller and power might be relatively low when the Type I error is well below the nominal level (Wilcox, 2001).

The contents of the thesis is for internal user only

REFERENCES

Abdullah, S. (2011). Kaedah Alexander – Govern dengan pendekatan pangkasan data: satu kajian simulasi. (Unpublished doctoral dissertation). Universiti Utara Malaysia.

Abdullah, S., Syed Yahaya, S. S., & Othman, A. R. (2010). Alexander-Govern test using adaptive trimmed mean as alternative to t-test.

Bluman, A. G. (2011). *Elementary statistics: a step by step approach* (8^{th}ed.). New York: McGraw-Hill.

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology*, 31, 144 – 152.

Cohen, J. (1976). *Statistical power analysis for the behavioral sciences* (*2nd ed*.). Hillsdale, NJ: Lawrence Erlbaum.

Efron, B. (1979). Bootstrap methods: another look at the Jackknife. *The Anneal of Statistics*. *7*(1): 1 – 26.

Erceg-Hurn, D. M., & Mirosevich, V. M. (2008). Modern robust statistical methods. *American Psychologist*, 591 – 601.

Geary, R. C. (1947). Testing for normality. *Biometrika*. 34, 209 – 242.

Hogg, R. V. (1974). Adaptive robust procedures. A partial review and some suggestions for future applications and theory. *Journal of the American Statistical Association*, 69, 909 – 927

Keselman, H. J., Cribbie, R. A. & Wilcox, R. (2002). Pairwise multiple comparison tests when data are nonnormal. *Educational and Psychological Measurement*, 62, 420-434.

Keselman, H. J., Othman, A. R., Wilcox, R. R. & Fradette, K. (2004). The new and improved two-sample t test. *Psychological Science*, *15*(1), 47.

Keselman, H. J., Wilcox, R. R., Algina, J. & Othman, A. R. (2004). A power comparison of robust test statistics based on adaptive estimators. *Journal of Modern Applied Statistical Methods*, *3*(1), 27 – 38.

Keselman, H. J., Wilcox, R. R., Lix, L. M., Algina, J. & Fradette, K. (2007). Adaptive robust estimation and testing. *British Journal of Mathematical and Statistical Psychology*, 60, 267 – 293.

Kowalchuk, R. K., Keselman, H. J., Wilcox, R. R. & Algina, J. (2006). Multiple comparison procedures, trimmed means and transformed statistics. *Journal of Modern Applied Statistical Methods*, 5, 43 – 64.

McClave, Benson & Sincich (2007). *Statistics for business and economics* (9th ed.). Singapore: Pearson Education South Asia Pte Ltd.

McHugh, M. L. (2011). Multiple comparison analysis testing in ANOVA. *Biochemia Medica*, *21*(3), 2011, 203 – 209.

Md. Yusof, Z., Othman, A. R. & Syed Yahaya, S. S. (2010). Comparison of Type I errorrates between T_1 and F_t statistics for unequal population variance using variable trimming. *Malaysian Journal of Mathematical Sciences*, *4*(2), 2010, 195 – 207.

Muhammad Di, N. F. (2013). The robustness of *H* Statistic with Hinge estimators as the location measures. (Unpublished doctoral dissertation). Universiti Utara Malaysia.

Muhammad Di, N. F., Syed Yahaya, S. S. & Abdullah, S. (2014). Comparing groups using robust *H* statistics with adaptive trimmed mean. *Sains Malaysiana*, *43*(4), 643 – 648.

Multiple-Comparison Procedures. Retrieved from http://www2.hawaii.edu/~taylor/z631/multcomp.pdf/

Newsom (2006). Post hoc tests. *USP 534 Data Analysis I Spring*. Retrieved from www.strath.ac.uklaer/materials/4dataanalysisineducationalresearch/unit6/post-hoctests/

Newsom (2012). Post hoc tests. *USP 634 Data Analysis I Spring*. Retrieved from http://web.pdx.edu/~g3jn/da1/ho_post%20hoc.pdf

Reed, J.F. & Stark, D. B. (1996). Hinge estimators of location: robust to asymmetry. *Computer Methods and Programs in Biomedicine*, 49, 11 – 17.

Rousseeuw, P. J. & Leroy, A. M. (2003). Robust regression and outlier detection. United States of America.

Singh, K. (1998). Breakdown theory for bootstrap quantiles. *Annals of Statistics*, 26, 1719 – 1732.

Spinella, S. (2011). Using the descriptive bootstrap to evaluate result replicability (because statistical significance doesn't). Texas A&M University.

Staudte, R. G. & Sheather, S. J. (1990). *Robust Estimation and Testing*. John Wiley & Sons Inc., New York.

Stuart, J. P., Nancy, L. G. & Anastasios, A. T. (1987). The analysis of multiple endpoints in clinical trials. *Biometrics*, *43*(3), 487 – 498.

Wilcox, R. R. (1997). Introduction to Robust Estimation and Hypothesis Testing. Academic Press, New York.

Wilcox, R. R. (2001). Pairwise comparisons of trimmed means for two or more groups. *Psychometrika*, *66*(3), 343 – 356.

Wilcox, R. R. & Keselman, H. J. (2002). Power analyses when comparing trimmed means. *Journal of Modern Applied Statistical Methods*, *1*(1): 24 – 31.

Wilcox, R. R. (2003). Multiple comparisons based on a modified one-step *M*-estimator. *Journal of Applied Statistics, 30*(10), 1231 – 1241.

Williams, L. J. & Abdi, H. (2010). Fisher's least significant difference (LSD) test. *Encyclopedia of Research Design*.

Zachary, R. S. & Craig, S. W. (2006). Central limit theorem and sample size. Retrieved from http://www.umass.edu/remp/Papers/Smith&Wells_NERA06.pdf