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-*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-*HQ_1) and *P*-Mean were compared with the existing LSD-Bonferroni correction. The finding revealed that *P-HQ* and *P-*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-*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).

Most literature stated that, if a population has a normal probability distribution, then for any sample sizes, the sampling distribution of the mean will also be normally distributed as concurred by Cohen (1976), McClave, Benson and Sincich (2007), and Bluman (2011), to name a few. Their statements emphasized that for sufficiently large samples, the central limit theorem will work irrespective of the population distribution shape. However, Geary (1947) stated that, "Normality could be viewed as a special case of many distributions rather than a universal property." He even advised that all existing textbooks as well as new textbooks should include this warning: "Normality is a myth; there never was, and never will be, a normal distribution" in future editions. Micceri (1989) as cited in Zachary and Craig (2006) analyzed 440 distributions from all different sources with sample size ranged from 190 to 10,893 found that no distributions among those investigated passed *Kolmogorov Smirnov* test of normality, and very few seem to be even reasonably close approximations to the Gaussian. In real life, the condition of normality and variance homogeneity is hardly attained (Erceg-Hurn & Mirosevich, 2008). Furthermore, there is no quantitative indication is given to indicate the degree of distortion, in any condition.

Using MCPs that are rigidly depending on such assumptions will cause the procedure to be unsatisfactory and suspicious. If parametric tests are used under violation of assumptions, the risk of falsely rejecting the null hypothesis will increase, thus affecting the validity of the results. For such reason, extra care is needed to handle outliers that cause non normality since they may severely affect the data analysis. One potential solution to this Type I error inflation and power deflation is by substituting the usual least squares estimators with estimators which are less influenced by the effects of non-normality. Robust statistic has been defined with the features of maintaining adequate Type I error and statistical power (Erceg-Hurn & Mirosevich, 2008). It is an alternative approach to the standard statistical methods, producing estimators which are not excessively affected by outliers, and having reasonable efficiency when assumptions are violated. Of late, alternative for the usual least squares estimator has been determined in most research, which is robust location measures such as trimmed means (Wilcox & Keselman, 2002; Keselman, Othman, Wilcox & Fradette, 2004; Kowalchuk, Keselman, Wilcox, & Algina, 2006; Md.Yusof, Othman & Syed Yahaya, 2010). However, to avoid the information loss during the trimming process, trimming need to be implemented carefully. For occasion when sampling from a light tailed distribution, very few observations are potentially to be trimmed, while if the sample is being taken from a normal distribution, trimming can be avoided. A normal response is that more observations will be trimmed from the right tail rather than from the left tail of the distribution for right skewed distribution. When using the usual trimmed means, the approach to reduce the distributions tails' effects is by solely eliminating them according to the predetermined amount. When this usual method is being applied, trimming will also be carried out to the observations from normal distribution according to the predetermined amount, for instance 10% or 20% on both tails. However, the trimming of the observations from a normal distribution can be excluded. To avoid unnecessary trimming and to trim accordingly, Keselman *et al.* (2007) applied asymmetric trimming. Specifically, hinge estimators recommended by Reed and Stark (1996) had been applied to define the appropriate trimming amount on each tail of a distribution. In this research, this estimator was known as adaptive trimmed mean and was proven to work well with the Welch test.

Reed and Stark (1996) has proposed adaptive trimmed mean using hinge estimator as one type of robust central tendency estimator. They produced seven hinge estimators which are HQ, HQ_1, HQ_2, HH_1, HH_3, HSK_2 and HSK_5. Among these estimators, the two best estimators recommended are HQ and HQ_1. The adaptive trimmed mean applies these two hinge estimator with the purpose to regulate the trimming process which outfits the data distribution shape. When these two estimators are used with trimmed means, they will have good control in Type I error rate. This was proven in Abdullah, Syed Yahaya and Othman (2010) when the two estimators were applied as central tendency measures in Alexander Govern test. These estimators were later tested on *H*-statistic. The result also showed good control of Type I error rates of the *H*-statistic. Nevertheless, testing on the groups is only possible for omnibus test, but could not continue with the MCP even if significant differences are detected due to the nonexistence of the MCP for these estimators. Since the distributions of the estimators as well as the test statistics are intractable, it will be inaccurate to use the default MCPs in most of the software.

This study proposed to develop a suitable MCP for HQ and HQ_1 using the method suggested by Wilcox (2003) known as *P*-Method. This method has resulted in good Type I error control when used on modified one-step M-estimator.

1.2 Problem Statement

There are quite a number of robust omnibus tests available to be used for testing the equality of groups. However, the post hoc tests (or MCPs) for these robust omnibus tests are very limited and these post hoc tests are very much depend on the robust estimator used in each omnibus test. For example, *MOM-H*, a robust omnibus test, uses modified one-step *M*-estimator (*MOM*) as its center measure. Thus, the post hoc test which corresponds to the omnibus test should be testing for the modified one-step *M*-estimator. The MCPs that are available in the common statistical packages such as LSD, Scheffé, Tukey, and Bonferroni are for testing the classical means. Besides *MOM*, another robust estimator which is gaining acceptance due to its ability to control the effect of non-normality is the adaptive trimmed mean. Adaptive trimmed mean comes in various types depending on the hinge estimator used. In this study, two adaptive trimmed means, *HQ* and HQ_1 were considered to be used as the center measures for post hoc tests due to their good performance in controlling Type I error when tested on the *H*-statistic (Muhammad Di, 2013; Muhammad Di, Syed Yahaya & Abdullah, 2014). Without any post hoc tests for *HQ* and HQ_1, any test on equality of groups using these estimators will be incomplete, and the good performance of these estimators on *H* statistic for example will not be fully used. Therefore, this study proposed to construct post hoc tests on *HQ* and HQ_1 using a method known as *P*-Method. Simultaneously, this study will also look at the performance of $\bar{x}$ on *P*-Method as an improvement for the existing post hoc tests for the usual mean.

1.3 Research Objective

There are 4 objectives need to be accomplished in this study.

i. Develop algorithm and programming for robust Multiple Pairwise Procedure (post hoc test) – *P*-Method.

ii. Evaluate the performance of the *P*-Method for *HQ*, HQ_1 and the classical mean in terms of Type I error.

iii. Compare the performance of the *P*-Method for classical mean with parametric LSD-Bonferroni correction method.

iv. Application of *P*-Method in real data.

1.4 Significance of Study

This study will contribute to the development of knowledge in experimental design methodology and to mention, especially in the experimental sciences. Statisticians are mindful that, experimental design methodology is very much depending on assumptions of normality and equal variances among treatment groups. Nevertheless, data are rarely normally distributed in the real world. With these new alternative methods, this research will bring benefit to the researches (in various fields, especially the experimental sciences). They will not be constrained or being controlled with all the normality and homogeneity of variances assumptions. They can instead work with the original data and not to worry about the distributions shape.

CHAPTER TWO

LITERATURE REVIEW

2.1 Introduction

Normality and variance homogeneity assumptions must be fulfilled when parametric tests such as *ANOVA* or *t*-test are used to test the equality of central tendency measures. However, real world data are mostly non-normally distributed in nature. The application of classical parametric test such as *ANOVA* when the assumptions are violated will cause the results of the test to be invalid (Erceg-Hurn & Mirosevich, 2008). Thus, robust statistics are developed to deal with problems which are related to violation of assumptions.

Robust statistical tests are developed with the features of maintaining adequate Type I error control and reasonable statistical power when the data are either normally or non-normally distributed. Type I error is in place when the null hypothesis is falsely rejected. Whereas the probability that a Type II error will not occur is defined as the power of a test. Rousseeuw and Leroy (2003) stated that robust statistics are estimators that are not easily affected by outliers. Basically, robust statistical tests work by replacing the traditional estimators with robust estimators. For instance, Keselman *et al*. (2007) proposed trimmed mean and Winsorized variances as the central tendency measure and scale estimator respectively in *Welch* test.

2.2 Trimming: A Robust Approach

Trimming is the most frequently used method when dealing with skewed data. The amount of trimming and trimming methodology are the main factors in trimming. There are two common trimming methods, which are symmetric trimming and asymmetric trimming. Equal amount of data will be trimmed away from both tails of the distributions in symmetric trimming. On the other hand, in asymmetric trimming, the trimming process is conducted on either one-tail or both tails with different amount. Trimming need to be processed carefully in order to avoid information loss. Usually, the amount of trimming has to be defined first before trimming being performed. Amount of trimming can be determined either by predetermined the trimming amount (fixing the trimming amount) or empirically specified using the trimming methods.

2.2.1 Usual Trimmed Mean

What is trimmed mean? Trimmed mean is an alternative to the mean and the median (Erceg-Hurn & Mirosevich, 2008). According to Wilcox (2001), the result will be different when the distribution slightly departs from normality towards heavy tail. Result has shown that 20% trimmed mean is more efficient than mean. However, trimming has to be done carefully to avoid the loss of information. At 20% trimming, 20% of the observations will be trimmed from both tails. The trimmed mean is

$$\overline{X_t} = \frac{1}{n-2g}\left(X_{(g+1)} + \cdots + X_{(n-g)}\right). \tag{2.1}$$

where observation proportion to be trimmed is γ. Quantity of observations to be trimmed from both tails is determined by $g = [\gamma n]$.

However, there are two drawbacks regarding trimmed mean (Wilcox, 2001). The first concern is related to symmetric trimming. Twenty percent trimmed mean may work well if the distribution is symmetry. But if the distribution is not symmetry, for example, a heavy right skewed distribution, intuition will suggest to trim more observations at right tail as compared to left tail. In other case, if it is a normal distribution, trimming might not be needed at all. The second concern is fixing the amount of trimming to be used. The common trimmed mean uses the predetermined method for trimming. By using this method, the pre-determine amount such as 10% or 20% will decide on how many observations should be trimmed equally on both tails of a distribution. For many situations, 20% trimmed mean seem adequate to trim away the noisy data, or in other word, to remove the outliers. However, problems arise when the percentage of outliers is greater than 20%. One might use the higher amount of trimming, but higher amount of trimming will lower the test power if the light tailed distribution is used, where outliers are relatively rare (Wilcox, 2001). Low breakdown point is another concern of trimmed mean. Trimmed mean has low breakdown point which relies on the percentage of trimming. Thus, to perform optimum trimming and to avoid unnecessary trimming, *MOM* (Wilcox, 2003) and hinge estimators (Reed & Stark, 1996) will be the good choice as the alternative for trimmed mean. *MOM* and hinge estimators employ asymmetric trimming. Suitable amount of trimming will be determined on each distribution tail which will minimize the risk of loss of useful data.

2.2.2 Adaptive Trimmed Means with hinge Estimators

Asymmetric trimmed mean uses hinge estimators to determine the data proportion to be trimmed from each tail based on the distribution shape. Hinge estimators are used by adaptive trimmed mean to adjust the trimming process to be suited to the data distribution. Breakdown point of adaptive trimmed means with hinge estimators is relied on the percentage of trimming.

There are seven location estimators defined by Reed and Stark (1996), which are Q, Q_1, Q_2, H_3, H_1, SK_2 and SK_5. These selector statistics are used to measure the tail length and skewness of distribution. HQ and HQ_1 have been proven to be the two best hinge estimators (Keselman, Wilcox, Lix, Algina & Fradette, 2007). The adaptive trimmed means using these two hinge estimators produce very good control of Type I error rates when applied in some robust procedures such as Welch test. Keselman *et al*. (2007) found that HQ and HQ_1 remain robust when variance is unequal while the other estimators stated above are not consistently robust across the different degrees and pairings of heterogeneity.

Hogg (1974) defined two measures of tail length as shown below;

$$Q = (U_{(0.05)} - L_{(0.05)})/(U_{(0.50)} - L_{(0.50)}) \tag{2.2}$$

$$Q_1 = (U_{(0.20)} - L_{(0.20)})/(U_{(0.50)} - L_{(0.50)}) \tag{2.3}$$

where

L_α = mean of the smallest αn observations

U_α = mean of the largest αn observations.

For example, if $\alpha = 0.02$, then $L_{(0.02)}$ is the mean of the smallest $0.02n$ observations.

These two selector statistics are applied to categorize the distributions type, such as light-tailed distribution, medium-tailed distribution or heavy-tailed distribution. Both Q and Q_1 are location free. Moreover, they are uncorrelated with location statistics in the event where trimmed means had. Table 2.1 and 2.2 below summarizes the Q and Q_1 tail length measurements respectively.

Table 2.1

Q Tail Length Measurement

Q	
***Q* values**	**Distributions**
$Q < 2.0$	Light-tailed (Uniform)
$2.0 < Q \leq 2.6$	Medium-tailed (Normal)
$2.6 < Q \leq 3.2$	Heavy-tailed
$Q > 3.2$	Very Heavy-tailed

Table 2.2

Q_1 Tail Length Measurement

$\boldsymbol{Q_1}$	
$\boldsymbol{Q_1}$ **values**	**Distributions**
$Q_1 < 1.81$	Light-tailed (Uniform)
$1.81 \leq Q_1 \leq 1.87$	Medium-tailed (Normal)
$Q_1 > 1.87$	Very Heavy-tailed

2.2.2.1 Hinge Estimator, *HQ*

Reed and Stark (1996) defined *HQ* as,

$$\propto_l = \propto \left[\frac{UW_Q}{UW_Q + LW_Q}\right] \tag{2.4}$$

$$HQ = \frac{UW_Q}{UW_Q + LW_Q} \tag{2.5}$$

where

$\propto = Total\ Amount\ Trimming\ from\ the\ Sample$

$\propto_l = Lower\ Trimming\ Proportion$

$\propto_u = Upper\ Trimming\ Proportion = \propto - \propto_l$

$$UW_Q = \left[\sum_J n_j (U_{0.05} - L_{0.05})\right] / \sum_j n_j \tag{2.6}$$

$$LW_Q = \left[\sum_J n_j (U_{0.5} - L_{0.5})\right] / \sum_j n_j . \tag{2.7}$$

The difference between *HQ* and HQ_1 is the calculation of L_α.

Keselman *et al*. (2007) stated that *HQ* has good control on Type I error protection, commits Type I error which only excess 10% from Bradley's upper bound of 0.075.

2.2.2.2 Hinge Estimator, HQ_1

Reed and Stark (1996) defined HQ_1 as,

$$\propto_l = \propto \left[\frac{UW_{Q1}}{UW_{Q1} + LW_{Q1}}\right] \tag{2.8}$$

$$HQ_1 = \frac{UW_{Q1}}{UW_{Q1} + LW_{Q1}} \tag{2.9}$$

where

$\propto = Total\ Amount\ Trimming\ from\ the\ Sample$

$\propto_l = Lower\ Trimming\ Proportion$

$\propto_u = Upper\ Trimming\ Proportion = \propto - \propto_l$

$$UW_{Q1} = \left[\sum_J n_j (U_{0.2} - L_{0.2})\right] / \sum_j n_j \qquad (2.10)$$

$$LW_{Q1} = \left[\sum_J n_j (U_{0.5} - L_{0.5})\right] / \sum_j n_j. \qquad (2.11)$$

A simulation study by Keselman *et al*. (2007) showed that when two and four groups were tested, HQ_1 has the best performance among the seven hinge estimators, producing fewest numbers of liberal values. Besides, HQ_1 is the best in controlling Type I error and in achieving power to detect effects as compared to HQ. They also discovered that HQ_1 commits Type I error which only in excess of 5% from Bradley's upper bound of 0.075. It could be recommended as the hinge estimator to adopt to determine whether and how data should be trimmed asymmetrically.

Until recently, the works on the hinge estimators such as HQ and HQ_1 focused on omnibus test. What if the test result comes out significant? As of now, no post hoc test (MCP) has been developed for these estimators yet. In this study, we took the challenge of developing a MCP for HQ and HQ_1, with the intention that the procedure could pave a way for other researchers in using the MCP for other robust measures as well.

2.3 Multiple Pairwise Comparison Procedures

As mentioned earlier, the pairwise MCPs for the test statistics using hinge estimator (Reed & Stark, 1996) have yet to be developed. This study is carried out with the objective to develop a suitable pairwise MCP for hinge estimator.

2.3.1 Percentile *t*-Bootstrap

Wilcox (1997) recommended percentile *t*-bootstrap when comparing 20% trimmed mean. In percentile *t*-bootstrap procedure, quantiles of a Studentized maximum modulus distribution will be estimated by applying bootstrap approach. However, simulation has shown that when sample sizes are small, the Type I error rates will drop well below the nominal level (Wilcox, 2001). *P*-Method, *PW*-Method and *PTW*-Method had been introduced as the alternatives to percentile *t*-bootstrap method (Wilcox, 2001).

2.3.2 Percentile Bootstrap Method, Method *P*

P-Method has advantages when dealing with contaminated data. As compared to percentile *t*-bootstrap, *P*-Method is found to have better control on Type I error rates which ensure that the probability of Type I error does not exceed the nominal level as well as does not drop too far-off below the nominal level (Wilcox, 2001). Simulation had shown that although the sample sizes are as small as 11, *P*-Method still has the ability to maintain the good control in Type I error rates. *P*-Method disadvantage is the relatively long confidence interval as compared to *PW*-Method. Nevertheless, Wilcox (2001) noted that when there are outliers, these two methods give very similar results.

2.3.3 Method *PW* and *PTW*

PW-Method performs reasonably well if the sample sizes are not too small. It performed well when using Singh's method (Singh, 1998) in terms of Type I error with the condition that sample is equal to or larger than 15 (Wilcox, 2001).

Simulation had shown that *PW*-Method can only control Type I error with a sample size of at least 15. Wilcox (2001) noted that *PW*-Method might give substantially shorter confidence interval as compared to *P*-Method, but the improvement is modest in most cases (increase in power is rather small). Moreover, it does not perform as well as *P*-Method in terms of controlling the Type I error rates (Wilcox, 2001).

Another method proposed by Wilcox (2001) is *PTW*-Method which does not compete well with *P*-Method in terms of the confidence interval. Moreover, it performs rather poor when testing by using four independent groups.

2.4 LSD with Bonferroni Correction

The earliest MCPs that are commonly used are LSD, Tukey, Scheffe and Bonferroni (McHugh, 2011). We will further discuss the details of each method in the following paragraph.

When the experimental and control group sizes are unequal, the Tukey method is recommended. Tukey test is easy to compute and it will test all pairwise differences. Moreover, it has the advantage of reducing the probability of making Type I error (McHugh, 2011). However, the Tukey method is not appropriate to test complex comparisons and it is not idyllic for exploratory studies. Besides, it is not as common as compared to Bonferroni and Scheffe.

Scheffe method is applicable for both simple and complex comparisons and it tests all possible comparisons. This is beneficial to the condition when we have no sufficient prior research to explain the findings. Retrieved 20th Dec 2015, from http://www2.hawaii.edu/~taylor/z631/multcomp.pdf, Scheffe is too conservative in the case when pairwise comparisons are the only comparison of interest. Besides, it is recommended for the tests when the consequences of Type II error be more important than the consequences of Type I error (McHugh, 2011). Scheffe requests the sample size to be equal as the condition of using this method (McHugh, 2011). The disadvantage of Scheffe is its demand of equal sample sizes.

Compared to Tukey method, Bonferroni has the advantage to test complex comparisons. When compared to Scheffe method on the other hand, Bonferroni has the advantage of controlling Type I error (McHugh, 2011). Bonferroni is available in many statistical packages. Bonferroni is known with its conservative character as well, that is the family wise error is less than alpha in many situations (Newsom, 2012). One of the disadvantages of Bonferroni is this method has to be applied only in the samples which have equal sizes and is not suggested for exploratory studies. It also limits the numbers of comparisons to be tested by requiring the researcher to specify all comparisons to be made in advance (McHugh, 2011). Therefore, less information on differences between the groups can be determined since not all differences are being tested.

LSD test stands for Least Significant Difference test. It was developed by Fisher in 1935 (Williams & Abdi, 2010). LSD is the oldest and one of its advantages is its simplicity. However, according to Newsom (2006), LSD is known with the ineffectiveness in controlling Type I error especially when uses on more than 3 groups. Due to the simplicity of LSD, in this study, we incorporated Bonferroni correction (Bonferroni, 1936, as cited in Stuart, Nancy & Anastasios, 1987) to the classical LSD to improve its family-wise error control. By using Bonferroni correction, α level of each individual test is adjusted downward. It sets the significant cut-off at α/n. The desired α value will be divided by the number of pairwise comparisons to be tested. For example, if the α value is set at 0.05 and the number of comparison to be conducted is 6, then the adjusted α value is now set as 0.0083, and with this, we still maintain 5% significance level in our analysis.

2.5 Bootstrap Method

Bootstrap method is popular in empirical research. It was introduced by Efron (1979) as a computer-based method for estimating the standard error of $\hat{\theta}$. In bootstrap, we are not only obtaining an estimate of the parameter but also to generate new samples. From here, many more estimators will be generated and hence an idea on estimator's variability obtained (Staude & Sheather, 1990). Bootstrap method is sampling within the sample. The sample is picked randomly from the data set and this selected number is then again replaced into the data and has the same chances to be drawn again. It treats the sample as population and draws samples from this pseudo population in order to assess

i. Variability of an estimator

ii. Bias of an estimator

iii. Predictive performance of a rule

iv. Significance of a test

Bootstrap has its advantages of wide application. In fact, it doesn't require theoretical calculation and is available no matter how mathematically complicated the estimator may be. Besides, it increases the estimator's accuracy. Last but not least, it has the ability to utilize the modern computing and it is completely automated. In the situation where the population data is not available, the values in the random sample is recognized as the best guide to the distribution, and resampling the sample is the best guide to what can be expected from resampling the population.

CHAPTER THREE
METHODOLOGY

3.1 Introduction

In this study, pairwise multiple comparison procedure for adaptive trimmed means (*HQ* and HQ_1) had been developed using the statistical test namely percentile bootstrap method (*P*-Method). *P*-Method has the advantage when dealing with contaminated data. It was found to have better control on Type I error rates as compared to the aforementioned MCPs, such that it neither exceed the nominal level nor drop too far below the nominal level (Wilcox, 2001). Figure 3.1 shows the procedures proposed in this study.

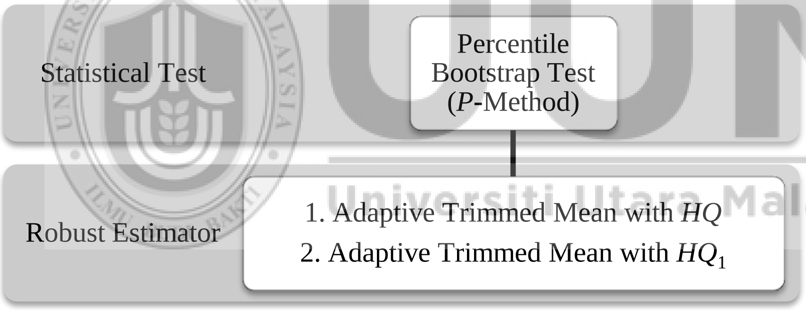

Statistical Test
Percentile Bootstrap Test (*P*-Method)
Robust Estimator
1. Adaptive Trimmed Mean with *HQ*
2. Adaptive Trimmed Mean with HQ_1

Figure 3.1. Statistical test with the corresponding robust estimators

3.2 *P*-Method

This study uses the percentile bootstrap method also known as *P*-Method as the alternative method for pairwise multiple comparison. The null hypothesis is,

$$H_0: \theta_j = \theta_k \tag{3.1}$$

for all $j < k$ and θ represents the parameter for either *HQ* or HQ_1.

Let X_{ij} such that i = 1, … , n_j and j = 1, … , J be a random sample of n_j observations from the jth group. Firstly, consider J = 2 and for fixed j let X_{1*j}, … , X_{n_j*j} be a bootstrap sample obtained by randomly resampling with replacement n_j observations from X_{1j}, … , $X_{n_j j}$. We let

$$P_{jk}^{*} = P\left(\hat{\theta}_j^{*} > \hat{\theta}_k^{*}\right). \tag{3.2}$$

where $\hat{\theta}_j^{*}$ is the value of $\hat{\theta}$ based on a bootstrap sample from the jth group.P_{jk}^{*} is the probability that the value of $\hat{\theta}_j^{*}$ is greater than $\hat{\theta}_k^{*}$ when resampling from the empirical distributions associated with the jth and kth groups. In other words, P_{jk}^{*} shows the degree to which the empirical distributions differ. $P*$ is equal to 0.5 if the empirical distributions are identical.

P_{jk}^{*} is estimated in the following manner. B bootstrap sample is generated from the jth group. Each bootstrap has n_i observations and $\hat{\theta}$ is computed for each bootstrap sample. The result is labeled as $\hat{\theta}_{jb}^{*}, b = 1, \dots, B$. Let I_b= 1 if $\hat{\theta}_{jb}^{*} > \hat{\theta}_{kb}^{*}$, otherwise, let I_b= 0. Then, estimate P_{jk}^{*}such that

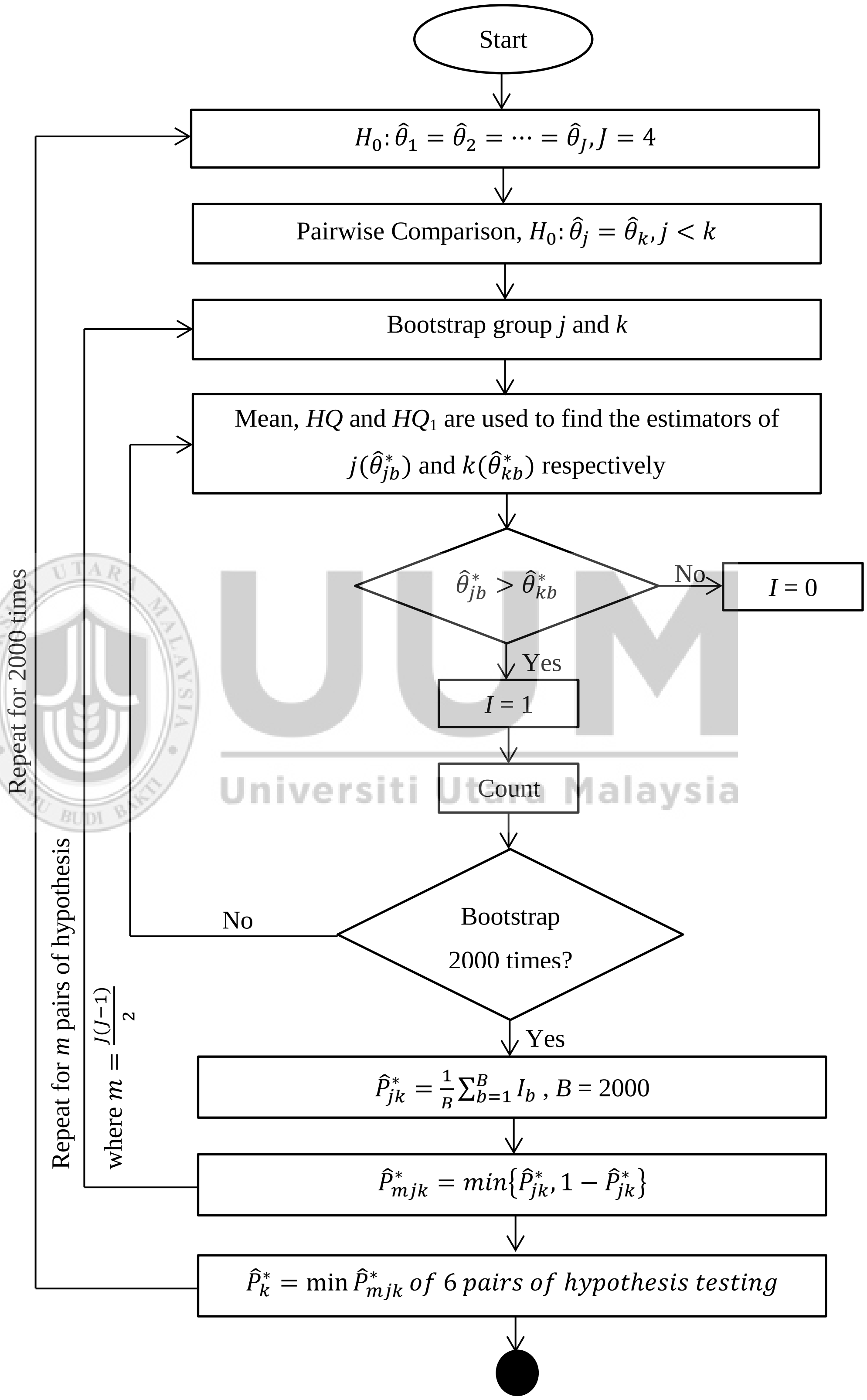
$$P_{jk}^{*} = \frac{1}{B}\sum_{b=1}^{B} I_b. \tag{3.3}$$

Let $\hat{P}^{*}$be the minimum value of $\hat{P}_{mjk}^{*}$over all j<k.

$$\hat{P}_{mjk}^{*} = min\{\hat{P}_{jk}^{*}, 1 - \hat{P}_{jk}^{*}\}. \tag{3.4}$$

$H_0: \theta_j = \theta_k$ is rejected if $\hat{P}_{mjk}^{*} < \propto_c$ and $\propto_c = 2 \times (1 - \Phi(q_c))$, Φ is the standard normal cumulative distribution and q_c is the 1 – α quantile of a Studentized maximum modulus distribution with a infinite degree of freedom. The estimation of α_c which is denoted as $\hat{\alpha}_c$ is based on 2,000 replications.

Figure 3.2 shows the algorithm for *P*-Method.


Start
$H_0: \hat{\theta}_1 = \hat{\theta}_2 = \cdots = \hat{\theta}_J, J = 4$
Pairwise Comparison, $H_0: \hat{\theta}_j = \hat{\theta}_k, j < k$
Bootstrap group j and k
Mean, HQ and HQ_1 are used to find the estimators of $j(\hat{\theta}^*_{jb})$ and $k(\hat{\theta}^*_{kb})$ respectively
$\hat{\theta}^*_{jb} > \hat{\theta}^*_{kb}$
No
$I = 0$
Yes
$I = 1$
Count
Bootstrap 2000 times?
No
Yes
$\hat{P}^*_{jk} = \frac{1}{B}\sum_{b=1}^{B} I_b$, $B = 2000$
$\hat{P}^*_{mjk} = min\{\hat{P}^*_{jk}, 1 - \hat{P}^*_{jk}\}$
$\hat{P}^*_k = \min \hat{P}^*_{mjk}$ $of\ 6\ pairs\ of\ hypothesis\ testing$
Repeat for 2000 times
Repeat for m pairs of hypothesis
where $m = \frac{J(J-1)}{2}$

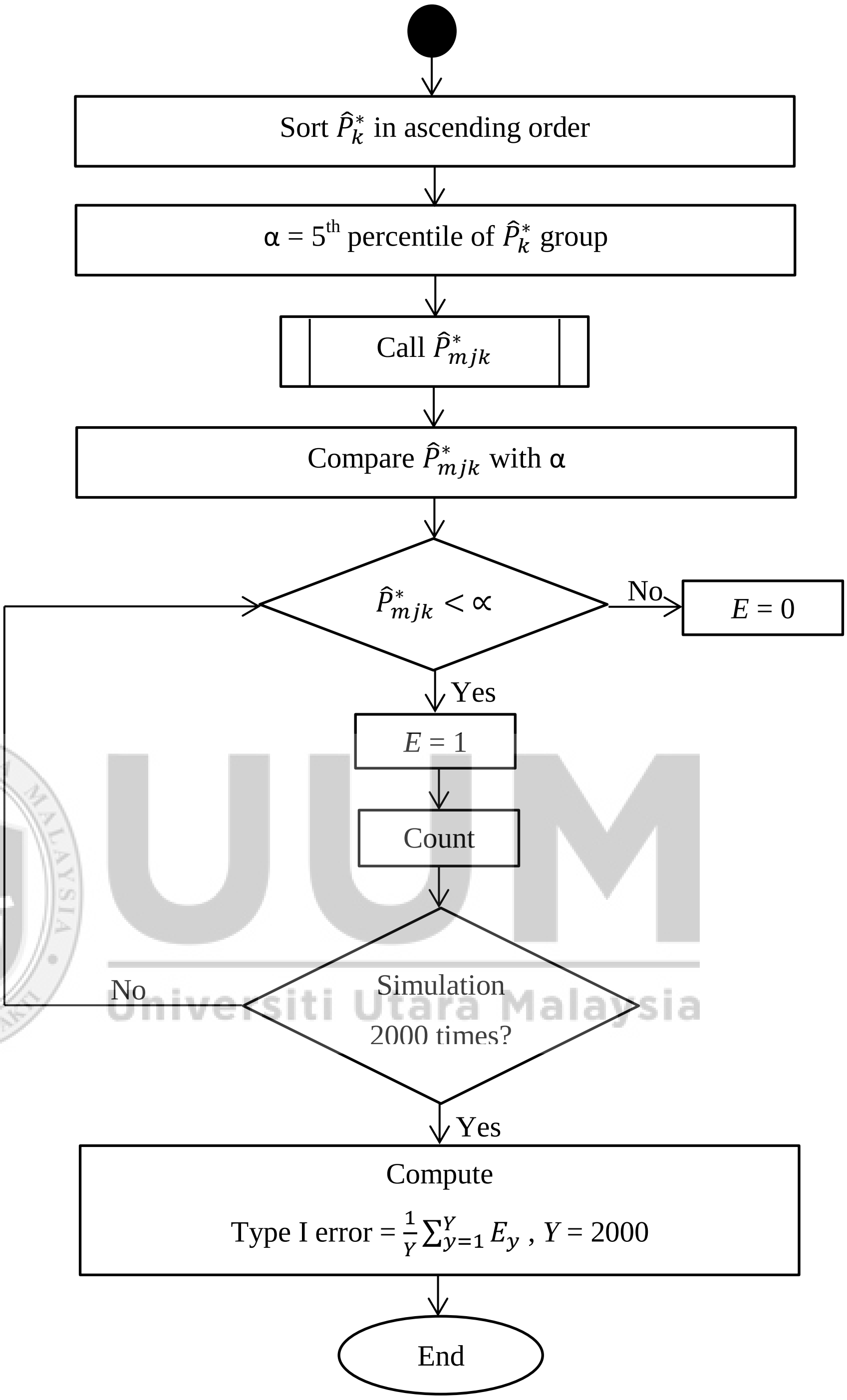

Sort $\hat{P}_k^*$ in ascending order
α = 5th percentile of $\hat{P}_k^*$ group
Call $\hat{P}_{mjk}^*$
Compare $\hat{P}_{mjk}^*$ with α
$\hat{P}_{mjk}^* < \propto$
No
$E = 0$
Yes
$E = 1$
Count
Simulation 2000 times?
No
Yes
Compute
Type I error = $\frac{1}{Y}\sum_{y=1}^{Y} E_y$, Y = 2000
End

Figure 3.2. P-Method algorithm

3.3 Variables Manipulated

To test the strength and weakness of the procedure, five variables as shown in Table 3.1 are manipulated.

Table 3.1

Descriptions of Variable Being Manipulated

No	Conditions	Descriptions
1	**Type of Population Distributions**	Normal Symmetric heavy tailed Skewed normal tailed Skewed heavy tailed
2	**Number of Groups**	$J = 4$ $J = 6$
3	**Sample Size**	Equal Unequal
4	**Degree/Pattern of Variance Heterogeneity**	Equal Moderate Large
5	**Pairing of Groups and Variances**	Positive Negative

Since these estimators need predetermined amount of trimming, the suitable percentage to be chosen is one of the main concerns. Keselman *et al*. (2007) and Abdullah (2011) recommended 15% total trimming to be used for HQ_1. By using 15% trimming, the power to detect HQ_1 effects is higher. Due to the recommendation, we adopt 15% trimming ($\alpha = 0.15$) in this study.

The quality of a statistical analysis is dramatically affected by real data which usually do not satisfy classical assumptions completely (Rousseeuw & Leroy, 2003). Nonetheless, the condition of normality and variance homogeneity is hardly attained in real life. Robust statistic enables the researchers to analyze the original data without worrying about the distributions shape. To investigate the effect of non-normality on the Type I error, four types of distributions are chosen. The distributions which represent normal and non-normal shapes are generated using the *g* and *h* distribution as shown in Table 3.2 below.

Table 3.2

g - and - h Distributions

Conditions	**Descriptions**	***g* - and - *h* Distributions**	
		g	***h***
Type of Population Distributions	**Normal**	0	0
	Symmetric heavy tailed	0	0.5
	Skewed normal tailed	0.5	0
	Skewed heavy tailed	0.5	0.5

Comparisons are done on randomized designs which contains four groups (J = 4) and six groups (J = 6). Equal (balance) and unequal (unbalance) sample sizes are assigned to each group. The total sample sizes for J = 4 is set at 80. For equal sample sizes the number of observations are distributed as $n_1 = n_2 = n_3 = n_4 = 20$, while for unequal sample sizes, different number of observations had been assigned to each group such that $n_1 = 10,\ n_2 = 15,\ n_3 = 25,\ n_4 = 30$. For J = 6, the total sample sizes is set at 120. The number of observations for equal and unequal sample sizes are $n_1 = n_2 = n_3 = n_4 = n_5 = n_6 = 20$ and $n_1 = 10, n_2 = 15, n_3 = 15, n_4 = 25, n_5 = 25, n_6 = 30$ respectively.

Homogeneity of the variances is another assumption which is always being violated. This study investigated on three degrees of variances i.e. equal variances, moderate and extreme degree of heterogeneity for the effect of unequal variances on Type I error. For *J* = 4, the equal variances is set at 1:1:1:1, and for moderate and extreme degree of heterogeneity, the unequal variances are 1:4:16:36 and 1:1:1:36 respectively. On the other hand, for *J* = 6, the ratio for equal variances is 1:1:1:1:1:1, while for moderate and extreme degree of heterogeneity, the ratio are set at 1:4:4:16:16:36 and 1:1:1:1:1:36 respectively.

Positive and negative pairings are formed when unequal variances are paired with unequal sample sizes. For positive pairing, largest number of group observations is paired with largest group variance, and at the same time, smallest group observations pair with smallest group variance. On the other hand, negative pairing involves the pairing of largest number of group observations with the smallest group variance, while smallest number of group observations is paired with the largest group variance.

The pairings for moderate and extreme degree of heterogeneity with respect to number of groups (*J* = 4 and *J* = 6) are exhibited in Table 3.3 to Table 3.6.

Table 3.3

Nature of Pairing with Moderate Degree of Heterogeneity for J = 4

Group Variances (Moderate)				
Group	**1**	**2**	**3**	**4**
Group Sizes	10	15	25	30
Positive Pairing	1	4	16	36
Negative Pairing	36	16	4	1

Table 3.4

Nature of Pairing with Extreme Degree of Heterogeneity for J = 4

Group Variances (Extreme)				
Group	**1**	**2**	**3**	**4**
Group Sizes	10	15	25	30
Positive Pairing	1	1	1	36
Negative Pairing	36	1	1	1

Table 3.5

Nature of Pairing with Moderate Degree of Heterogeneity for J = 6

Group Variances (Moderate)						
Group	**1**	**2**	**3**	**4**	**5**	**6**
Group Sizes	10	15	15	25	25	30
Positive Pairing	1	4	4	16	16	36
Negative Pairing	36	16	16	4	4	1

Table 3.6

Nature of Pairing with Extreme Degree of Heterogeneity for J = 6

Group Variances (Extreme)						
Group	**1**	**2**	**3**	**4**	**5**	**6**
Group Sizes	10	15	15	25	25	30
Positive Pairing	1	1	1	1	1	36
Negative Pairing	36	1	1	1	1	1

In general, three types of conditions had been used in this study.

i. Perfect condition: the condition with normal distribution, equal group sizes and equal variances.

ii. Moderate condition:

a. Skewed distribution with equal sample sizes and equal variances.

b. Normal distributions with unequal sample sizes and unequal variances.

iii. Extreme condition: skewed distributions with unequal sample sizes and variances.

The design of the study is summarized in Figure 3.3

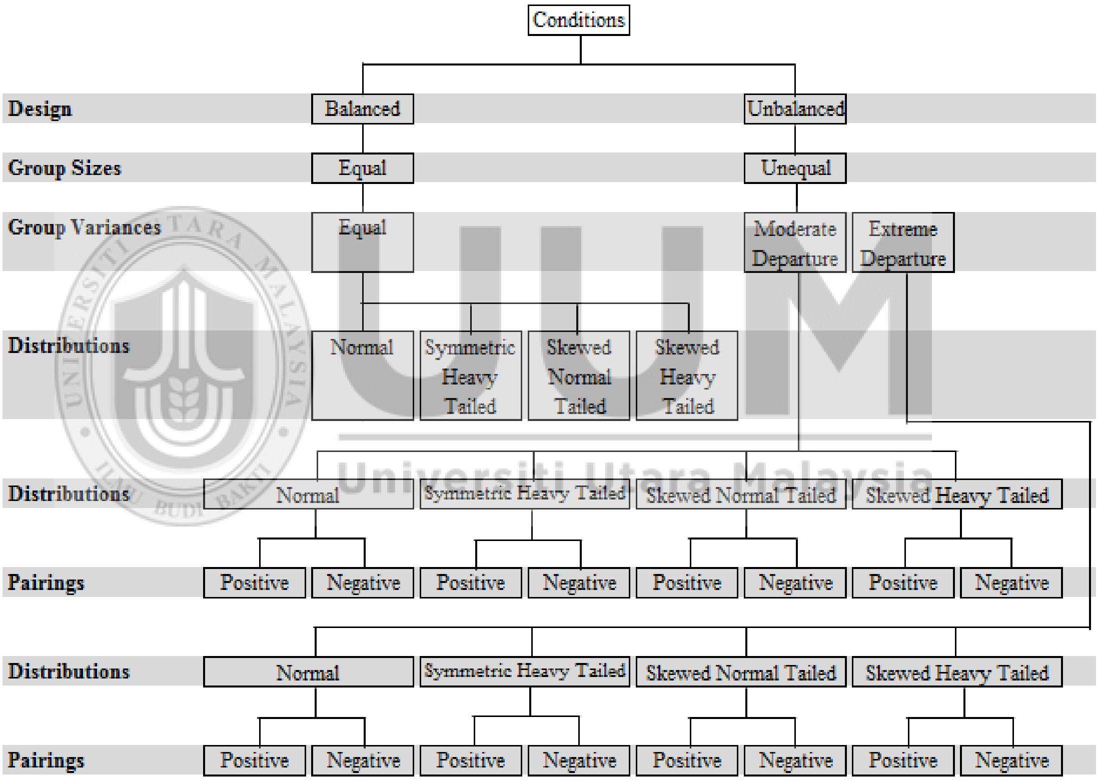


Figure 3.3. Summary of variances manipulated in this study

3.4 Data Generation

Type I error for each test is determined using 2,000 simulated datasets which had been generated using *SAS/IML* Version 9.2 (2008). The investigation on distribution shape is based on the types of distribution as explained below:

i. Normal distribution

We apply straight forward usage of SAS generator RANNOR with mean 0 and standard deviation 1 in our study to create standard normal distribution.

ii. Skewed heavy tailed (g = 0.5, h = 0.5) distribution

Standard normal distribution has been generated as in (i) and we transform it into random variables via equation below,

$$Y_{ij} = \frac{\exp(gZ_{ij})-1}{g}\exp(hZ_{ij}^2/2). \tag{3.5}$$

Skewness of the distribution is controlled by parameter g while kurtosis is controlled by parameter h. Tails of the distribution will become heavier as h increases while the degree of skewness will increase with the increasing of g. Here, we have chosen g=h=0.5 to create extreme non-normal condition.

iii. Symmetric heavy tailed (g = 0, h = 0.5) distribution

Equation 3.5 has been modified as below for symmetric heavy tailed distribution.

$$Y_{ij} = Z_{ij}\exp(hZ_{ij}^2/2). \tag{3.6}$$

iv. Skewed normal tailed (g = 0.5, h = 0) distribution

For skewed normal tailed distribution, we have modified Equation 3.5 such that,

$$Y_{ij} = \frac{\exp(gZ_{ij})-1}{g}. \tag{3.7}$$

The equation mentioned above well explained on how we create conditions for moderate and extreme departure from normality by applying g-and-h distribution. In general, the central measures have values unequal to zero when dealing with skewed or non-normal distribution. To ensure the null hypothesis remains true, the observations, Y_{ij} from each simulated skewed distributions are shifted by subtracting the population central tendency parameter from the observations as equation stated below.

$$X_{ij} = Y_{ij} - \theta. \tag{3.8}$$

We have computed $\hat{\theta}$ with one million observations generated from the distribution under study in order to determine the values of θ (Keselman, Wilcox, Algina & Othman, 2004; Wilcox, 2003). Based on the one million observations, the population location parameters for all proposed procedures have been tabulated in Table 3.7.

Table 3.7

Location Parameters with Respect to Distribution

Distributions	**Mean**	***HQ***	***HQ*₁**
Normal	0	0	0
g = 0.5, *h* = 0	0.2653	0.2310	0.1910
g = *h* = 0.5	0.7895	0.5240	0.3480

Next, we will share some examples of data generation using SAS/IML for 15% trimming with *P-HQ*₁ procedure.

i. Normal distribution

```
MTEMP = RANNOR(J(N,1,SSEED));
```

```
YTEMP = MTEMP[1:N];
```

ii. Skewed heavy tailed (g = 0.5, h = 0.5) distribution

```
YTEMP1 = (EXP(TEMP#0.5)-1.0)/0.5#EXP(TEMP##2#0.5/2);
YTEMP = YTEMP1 - 0.348;
```

iii. Skewed normal tailed (g = 0.5, h = 0) distribution

```
YTEMP1 = (EXP(TEMP#0.5)-1.0)/0.5#EXP(TEMP##2#0/2);
YTEMP = YTEMP1 - 0.191;
```

The significance level for all the tests is set at α = 0.05. For each design, 2000 datasets are simulated and each simulated dataset are bootstrapped for 2000 times (refer Section 3.5). *P*-Method is used to test the hypothesis since the distributions of the test statistics are intractable. This study then continues to investigate on the effectiveness of the proposed method using real data on students' academic achievement.

3.5 Application of Bootstrap Method

In this study, bootstrap method is used to test the hypothesis. To obtain the *p*-value of the Mean, *HQ* and HQ_1 statistic by using the percentile bootstrap method, the following steps are adhered to (Wilcox, 1997).

i. Calculate Mean, *HQ* and HQ_1 based on the available data respectively.

ii. Generate bootstrap samples by randomly sampling with replacement n_j observations from the *j*th group yielding $Y^*_{1j}, Y^*_{2j}, \ldots, Y^*_{n_j j}$.

iii. Each of the sample points in the bootstrapped groups must be centered at their respective estimated medians so that the sample median is zero, such that

$C_{ij}^* = Y_{ij}^* - \widehat{M}_j$, i = 1, 2, …, n_j. The empirical distributions are shifted so that the null hypothesis of the equal medians among the *J* distributions is true. The strategy behind the bootstrap is to use the shifted empirical distributions to estimate an appropriate critical value.

iv. Let P^* be the value of *P* test based on the C_{ij}^* values.

v. Repeat Step ii to Step iv *B* times yielding $\hat{P}_{k_{11}}^*$, $\hat{P}_{k_{12}}^*$, …, $\hat{P}_{k_{1B}}^*$.

vi. Get *p*-value by choosing the minimum value from [(*k*-1) x 2] pairs of comparisons. All together there are *B* *p*-values.

vii. Get α by sorting *p*-values from smallest to largest and get the 5th percentile as the α value. Reject H_0 if *p*-value is smaller than α value.

viii. Calculate Type I error by dividing the number of rejections out of the number of simulation.

$$Type\ I\ error = \frac{Number\ of\ rejections}{Number\ of\ simulation}. \qquad (3.9)$$

Typically, 1000 to 5000 bootstrap samples are recommended (Spinella, 2011). In our study, each simulated dataset are bootstrapped for 2000 times. Wilcox (2001 & 2003) mentioned that when *J* = 2, with α = 0.05, increasing of *B* = 500 to 2000 doesn't help in Type I errors robustness. However, when *J* = 4, *B* = 2000 will have the advantage and improve the Type I error rates. According to Wilcox (2001 & 2003), large value of *B* is required as *J* increases. With the help of modern technology, advance computer speed has made this task possible.

3.6 Measure of Robustness

To determine the robustness of a procedure, the empirical Type I error rates are benchmarked using Bradley’s (1978) liberal criterion of robustness. According to this criterion, a test is considered robust when it’s Type I error (ά) is in the interval of 0.5α ≤ ά ≤ 1.5α. Thus, when α = 0.05, a test is considered robust in a particular condition if its’ empirical Type I error rate falls within the interval of 0.025 ≤ ά ≤ 0.075.

CHAPTER FOUR

RESULTS OF THE ANALYSIS

4.1 Introduction

The proposed pairwise multiple comparison procedure (MCP), *P*-Method, is employed to test two location estimators namely *HQ* and HQ_1. Each MCP is compared for their robustness in terms of Type I error. Various conditions such as shapes of distributions, equal or unequal group variances, balanced and unbalanced sample sizes, nature of pairings of sample sizes and group variances, which are known to highlight the strengths and weaknesses are used to test the procedures. These conditions are then organized into two types of designs known as balanced and unbalanced design based on the sample sizes and group variances. The procedures for each particular design are then tested with four ($J = 4$) and six ($J = 6$) groups. In next session, the results in the form of Type I error rates are tabulated.

Bradley's (1978) liberal criterion of robustness is used to evaluate and determine which conditions are insensitive to the assumption violations. According to this criterion, a test is considered robust when it's Type I error (ά) is in the interval of $0.5\alpha \leq ά \leq 1.5\alpha$. Thus, when $\alpha = 0.05$, a test is considered robust in a particular condition if its empirical rate of Type I error fall within the interval of $0.025 \leq ά \leq 0.075$. Correspondingly, a test is considered to be non-robust if its Type I error is out of this interval. A test which can produce Type I error rate closest to the nominal level is considered the best. On a more stringent checks on Type I error control, another criterion which is based on a narrower interval of 0.9α ≤ ά ≤ 1.1α could also be adopted. For α = 0.05, the interval will be in the range of 0.045 ≤ ά ≤ 0.055.

4.2 *P*-Method

The proposed procedure, the *P*-Method is put to test using two location estimators namely *HQ* and HQ_1. The goal of the *P*-Method is to test the pairwise comparison among groups when the omnibus test is significant.

The hypothesis for omnibus test is

$$H_0: \theta_1 = \theta_2 = \cdots = \theta_J. \qquad (4.1)$$

where θ is the location parameter to be tested.

While the hypothesis for pairwise comparison is

$$H_0: \theta_i = \theta_j \text{ for } i = 1, \ldots, n_j \text{ and } j = 1, \ldots, J, i < j \qquad (4.2)$$

Type I error for the *P*-Method for various conditions is calculated and the results are discussed in the following section.

4.2.1 Type I error for *J* = 4

Type I error results under various condition for *J* = 4 are presented in Table 4.1 and Table 4.2. These tables represent the performance of *P*-Method for unbalanced and balanced design respectively.

4.2.1.1 Unbalanced Design (*J* = 4)

The combination of unequal sample sizes and heterogeneous group variances generate various conditions categorized under unbalanced design. The results for the four groups of unequal sample sizes and unequal group variances are discussed in this section focusing on degree of heteroscedasticity and nature of pairings, with the purpose to identify under which conditions the *P*-Method is robust or able to control Type I error. Conditions with error rates within 0.025 to 0.075 intervals are considered robust.

Table 4.1

Type I error Rates for J = 4 under Moderate and Extreme Degree of Heteroscedasticity

Distribution	**Variance**	**Pairing**	***P*-Method with Corresponding Scale Estimators**		
			Mean	***HQ***	***HQ_1***
Normal ***g* = 0, *h* = 0**	Moderate	Positive	0.05800	**0.04900**	**0.05400**
		Negative	0.04150	0.03900	0.04200
	Extreme	Positive	0.06950	**0.04550**	**0.04550**
		Negative	0.05800	0.05700	**0.04850**
Symmetric heavy tailed ***g* = 0, *h* = 0.5**	Moderate	Positive	0.05800	**0.05200**	0.05650
		Negative	0.03650	**0.04650**	**0.05000**
	Extreme	Positive	0.04450	**0.04500**	**0.05400**
		Negative	**0.04850**	**0.04900**	**0.05100**
Skewed normal tailed ***g* = 0.5, *h* = 0**	Moderate	Positive	0.06000	0.05950	0.05750
		Negative	0.06850	0.04450	0.04150
	Extreme	Positive	0.06050	0.06200	**0.05250**
		Negative	0.07050	**0.05500**	0.05550
Skewed heavy tailed ***g* = 0.5, *h* = 0.5**	Moderate	Positive	0.06000	**0.05100**	0.05750
		Negative	**0.04750**	**0.04600**	**0.05450**
	Extreme	Positive	0.06000	0.04400	0.05900
		Negative	**0.05250**	**0.05100**	**0.05050**
Grand Average			0.05588	**0.04975**	**0.05188**

Indication:

▨	Type I error rates fulfill the Bradley's robust range
Bold	Type I error rates fulfill the robust stringent criterion

Across Table 4.1, we could observe that all the Type I error rates fulfill Bradley's criterion. Overall, 46% of the Type I error rates are within the stringent robust range.

In Table 4.1, the "Grand Average" value in the last row represents the overall performance of the *P*-Method across different distributions with respect to different location estimators namely the usual mean, *HQ* and HQ_1. The result of the *P*-Method based on the grand average values indicates that, regardless of the location estimators, the proposed method is very much in control of its Type I error rate whereby the rates are incredibly close to the nominal level of 0.05. The values for *P-HQ* and *P*-HQ_1 are 0.04975 and 0.05188 respectively and fall under the stringent criterion of robustness. Oppositely, Type I error rate generated by *P*-Mean is slightly away from the nominal level, which is 0.00587 away from 0.05. Generally, if we are to rank the performance of each procedure according to their gap to nominal level, *P-HQ* outperformed *P*-HQ_1 and followed by *P*-Mean.

Next, we will discuss in detail on the performance of *P*-Method on the three estimators with respect to different types of distributions. Firstly, when the level of skewness as well as kurtosis is zero, or in other word, under normal distribution, *P*-Method with all the three estimators, Mean, *HQ* and HQ_1 produce robust Type I error rates in the range of 0.03900 to 0.06950. HQ_1 control Type I error rates most stringently, with 75% (3/4) Type I error rates fulfill the stringent criterion of robustness under combination of both moderate and extreme variance with positive and negative pairing, followed by *HQ* and Mean.

For the symmetric heavy tailed distribution, *P*-Method across the 3 estimators is able to control Type I error rates effectively under both moderate and extreme variance for both positive and negative pairings. These three procedures manage to generate Type I error rates between 0.03650 to 0.05800. *HQ* performs the best, with all Type I error rates meet the stringent criterion range, followed by HQ_1 and Mean.

The next distribution is skewed normal tailed, whereby we can observe that both *HQ* and HQ_1 are equally good in controlling Type I error rates. They outperform Mean across moderate and extreme variances as well as positive and negative pairings. The Type I error rates generated are in the range of 0.04350 and 0.06450. Both *HQ* and HQ_1 have success rate of 25% (1/4) to produce stringent Type I error rates respectively. Meanwhile, Mean achieves 0% (0/4) of stringent Type I error rates.

Last but not least, for skewed heavy tailed distribution, the *P*-Method still perform well for all the estimators even under extreme condition. Here, *HQ* has the best performance, producing 70% (3/4) stringent Type I error rates, followed by Mean and HQ_1 which have equally good performance, records 50% (2/4) stringent Type I error rates.

To recapitulate, all the Type I error rates generated by *P*-Method with Mean, *HQ* and HQ_1 successfully met the Bradley's robust criterion. In general, from the analysis, we can conclude that, the *P*-Method can effectively controlled Type I error rates for the three estimators, Mean, *HQ* and HQ_1 regardless of the shapes of the distributions. The best performance in controlling Type I error rates goes to *HQ* (63% stringent Type I error rate), closely followed by HQ_1 (56% stringent Type I error rate) and

Mean (19% stringent Type I error rate). This is proven by the results which do not only satisfy the Bradley's robust criterion, but also the stringent criterion of robustness.

4.2.1.2 Balanced Design (*J* = 4)

Balance design is conducted on groups with equal number of observations and homogeneous group variances.

Table 4.2

Type I error Rates for J = 4 Balanced Design

Distribution	***P*-Method with Corresponding Scale Estimators**		
	Mean	***HQ***	***HQ*$_1$**
Normal ***g* = 0, *h* = 0**	0.06050	0.06300	**0.05400**
Symmetric heavy tailed ***g* = 0, *h* = 0.5**	0.06350	0.06700	0.05700
Skewed normal tailed ***g* = 0.5, *h* = 0**	0.05850	0.05800	0.07100
Skewed heavy tailed ***g* = 0.5, *h* = 0.5**	0.05900	0.06750	**0.05500**
Grand Average	0.06038	0.06388	0.05925

Indication:

▨	Type I error rates fulfill the Bradley's robust range
Bold	Type I error rates fulfill the robust stringent criterion

Results tabulated in Table 4.3 shows that all the Type I error rates for the Grand Average generated fall within the Bradley's robust interval, indicating that on average, the *P*-Method is robust when used on the estimators. HQ_1 (0.05925) produced the best result, followed by Mean (0.06038) and *HQ* (0.06388).

Under normal distribution, *P*-Method produces the best result with HQ_1 (0.054) surpasses the Mean (0.0605) and *HQ* (0.063). For symmetric heavy tailed distribution, the ranking of the estimators is similar to normal distribution with HQ_1 (0.057) still the best, next is the Mean (0.0635) and lastly is the *HQ* (0.067). When the distribution is skewed normal tailed, *HQ* (0.058) outperforms the Mean (0.0585) and HQ_1 (0.071). The Type I error rate for HQ_1 inflates quite badly approaching the upper limit of Bradley's interval. Next, under skewed heavy tailed distribution, HQ_1 (0.055) procedure produces Type I error closest to the nominal value, followed by Mean (0.059) and *HQ* (0.0675).

Overall, for balanced design, the *P*-Method displays good control of Type I error rates across all types of distribution. *P*-Method with HQ_1 deems to be the best under normal distribution, symmetric heavy tailed distribution and skewed heavy tailed distribution in terms of robustness, while *P*-Method with *HQ* is the most robust under skewed normal tailed distribution.

Comparison between the results of balanced and unbalanced designs for J = 4 case shows that the *P*-Method with Mean, *HQ* and HQ_1 proven to have better control of Type I error rates under unbalanced design as compared to balanced design. All the Type I error rates under unbalanced design satisfied the stringent criterion of robustness while all the rates under balanced design could only satisfy the Bradley's liberal criterion of robustness.

In a nut shell, this study showed that *P*-Method with Mean, *HQ* and HQ_1 has good control of Type I error across all types of distributions, be it balanced or unbalanced designs for *J* = 4 case.

4.2.2 Type I error for *J* = 6

Performance of *P*-Method in controlling Type I error under various condition for *J* = 6 are shown in Table 4.3 and Table 4.4 for balanced and unbalanced design respectively.

4.2.2.1 Unbalanced Design (*J* = 6)

Like the other cases, for the six groups, the presentation of the results is divided into degree of heteroscedasticity and nature of pairings. For this case, six groups of unequal sample sizes and unequal group variances are used to obtain the Type I error rates in Table 4.3.

Table 4.3

Type I error Rates for J = 6 under Moderate and Extreme Degree of Heteroscedasticity

Distribution	Variance	Pairing	***P*-Method with Corresponding Scale Estimators**		
			Mean	***HQ***	***HQ*$_1$**
Normal ***g* = 0, *h* = 0**	Moderate	Positive	**0.04580**	**0.05200**	0.06200
		Negative	0.06000	**0.05250**	**0.05250**
	Extreme	Positive	**0.05000**	0.08400	**0.05450**
		Negative	0.06000	0.07800	0.06050
Symmetric heavy tailed ***g* = 0, *h* = 0.5**	Moderate	Positive	**0.05350**	**0.05300**	0.07600
		Negative	0.05800	0.05700	0.05700
	Extreme	Positive	**0.04800**	0.05800	0.08250
		Negative	0.06300	0.05950	0.05750
Skewed normal tailed ***g* = 0.5, *h* = 0**	Moderate	Positive	0.06650	0.06450	0.06450
		Negative	**0.05350**	0.05750	0.06400
	Extreme	Positive	**0.04900**	**0.04750**	**0.05350**
		Negative	0.07350	0.06300	0.05800
Skewed heavy tailed ***g* = 0.5, *h* = 0.5**	Moderate	Positive	0.06000	0.06700	0.06950
		Negative	**0.05300**	0.06600	**0.05450**
	Extreme	Positive	0.05700	0.06250	0.04400
		Negative	0.06300	**0.05450**	**0.04650**
Grand Average			0.05711	0.06103	0.05981

Indication:

▨	Type I error rates fulfill the Bradley's robust range
Bold	Type I error rates fulfill the robust stringent criterion

The rates which are out of the Bradley's robust interval are un-highlighted. From Table 4.3, it is recorded that only 4 Type I error rates fail to fulfill Bradley's robust range, which is equal to 8% from the total Type I error rates. The "Grand Average" value in the last row of the table represents the overall performance of the *P*-Method across different distributions corresponding to each estimator. From table 4.3, we can observe that the Type I error rate performance for *P*-Method is the best with Mean

(0.00711 away from nominal level), followed by HQ_1 (0.00981 away from nominal level) and *HQ* (0.01103 away from nominal level). The details of the performance of *P*-Method with respect to the three estimators are elaborated next.

In the following paragraph, we will discuss on the performance of these three estimators across different types of distribution. First distribution that we will look at is normal distribution. Mean and HQ_1 performed equally well, with all the Type I error rates fall within the Bradley's robust interval regardless of the degree of heteroscedasticity and pairing types. Moreover, they manage to generate 50% (2/4) of Type I error rates which fall in the stringent robust range. On the other hand, *HQ* has good control of Type I error under moderate degree of variance heteroscedasticity for both positive and negative pairing. However, when it is under extreme variance heteroscedasticity, it produces high Type I error rates, which is 0.08400 and 0.07800 for positive and negative pairings respectively.

For the symmetric heavy tailed distribution, Mean and *HQ* show their effectiveness in controlling Type I error rates regardless the degree of heteroscedasticity and pairing by producing Type I error rates in the range of 0.04800 to 0.06300. However, Mean still outperforms *HQ* as it able to generate one extra Type I error rate which met the stringent robust criterion compared to *HQ*. Meanwhile HQ_1, manage to retain its robustness under negative pairing only for both moderate and extreme degree of heteroscedasticity. It performs badly under positive pairing.

The next distribution in our investigation is skewed normal tailed distribution. All three estimators still remain their robustness in controlling Type I error rates regardless of the degree of variance and type of pairings. As can be observed from Table 4.3, Mean manages to achieve 50% (2/4) of stringent Type I error rates, followed by *HQ* and *HQ*$_1$, with both producing 25% (1/4) of Type I error rates which met the stringent robust criterion range. Therefore, all estimators generate stringent Type I error rates when it is under extreme degree of heteroscedasticity and positive pairing.

The last distribution in our investigation which represents the extreme deviation from normality is skewed heavy tailed distribution. Under this distribution, all three estimators again retained their robustness in controlling Type I error rates. The rates generated are in the range of 0.04400 to 0.06950. Both *HQ* and *HQ*$_1$ have equally good performance, producing 50% (2/4) stringent Type I error rates, followed by Mean (25% of stringent Type I error rates).

With respect to the stability in the performance, that is the ability in controlling Type I error across all types of distributions, we would recommend *P*-Mean as the best procedure in six groups unbalanced design. This is because Mean is the only estimator which remain robust across various types of distributions, while *HQ* and *HQ$_1$* fail to control Type I error rates under normal distribution (both positive and negative pairing under extreme degree of heteroscedasticity) and symmetric heavy tail distribution (positive pairing for both moderate and extreme degree of heteroscedasticity).

4.2.2.2 Balanced Design (*J* = 6)

In this section, the performance of the *P*-Method with the estimators for six groups balanced design will be thoroughly discussed. Balanced design is the situation where the equal sample sizes and the homogenous group variances applied.

Table 4.4

Type I error Rates for J = 6 Balanced Design

Distribution	***P*-Method with Corresponding Scale Estimators**		
	Mean	***HQ***	***HQ_1***
Normal ***g* = 0, *h* = 0**	0.07550	**0.05500**	0.05700
Symmetric heavy tailed ***g* = 0, *h* = 0.5**	0.05700	0.06050	0.07050
Skewed normal tailed ***g* = 0.5, *h* = 0**	0.07100	0.05750	0.06100
Skewed heavy tailed ***g* = 0.5, *h* = 0.5**	0.05800	**0.04600**	0.06100
Grand Average	0.06538	**0.05475**	0.06238

Indication:

▨	Type I error rates fulfill the Bradley's robust range
Bold	Type I error rates fulfill the robust stringent criterion

Table 4.4 showed the Type I error rates of the method with Mean, *HQ* and HQ_1 as the central measures. Un-highlighted cell indicated the Type I error which is out of the Bradley's robust interval.

The overall performance of the method corresponding to Mean, *HQ* and *HQ_1*, is represented by “Grand Average” in the last row of the table. The results shows that all Type I error generated were within the Bradley’s robust interval which signify that all procedures are robust regardless of the distributional shapes. *HQ* (0.05475) produces the best result, followed by *HQ_1* (0.06238) and the Mean (0.06538). Nevertheless, as we browse across Table 4.4, we can see an un-highlighted cell under the mean column with Type I error rate slightly beyond (0.07550) the Bradley’s robust interval. Ironically, this cell corresponds to normal distribution, which indicates that the *P*-Method could not control Type I error as expected under perfect condition (normal and equal variances). For this condition *HQ* (0.055) generates Type I error closest to the nominal value while *HQ_1* (0.057) differs by plus 0.002. In contrast, for symmetric heavy tailed distribution, mean (0.057) performs the best, followed by *HQ* (0.0605) and *HQ_1* (0.0705). Meanwhile, under skewed normal tailed distribution, *HQ* (0.0575) produces Type I error rate closest to the nominal value while mean (0.071) ranked the last after *HQ_1* (0.061). Lastly, for skewed heavy tailed distribution, *HQ* (0.046) again generates the best result, followed by Mean (0.058) and *HQ_1* (0.061). Better rates are observed for skewed heavy tailed as compared to skewed normal tailed.

Overall, the *P*-Method shows good control of Type I error rates across all types of distribution under balanced design. *HQ* is the best under normal distribution, skewed normal tailed distribution and skewed heavy tailed distribution, while mean is the most effective under symmetric heavy tailed distribution.

When comparison is made between the results of balanced and unbalanced designs for J = 6 case, we can conclude that the *P*-Method had better control of Type I error rates under balanced design as compared to unbalanced design. All the Type I error rates under balanced design are within the Bradley's robust interval, while under unbalanced design, only Mean fulfilled the Bradley's robust interval for all conditions.

In conclusion, this study shows that *P*-Method with mean, *HQ* and HQ_1 are able to control Type I error well across all types of distributions for balanced design of J = 4 case, while for unbalanced design, the only estimator which had good control of Type I error rates is the Mean.

4.3 *P*-Method with Mean (*P*-Mean) versus LSD with Bonferroni Correction (LSD-Bonferroni Correction)

From the results discussed in previous sections, pairing of *P*-Method and classical Mean has shown the capability to control Type I error rates effectively across all types of distributions for both J = 4 and J = 6. In the next section, we will compare and discuss on the performance of LSD-Bonferroni correction with *P*-Mean.

4.3.1 Type I error for *J* = 4

Type I error results for *P*-Mean and LSD-Bonfferoni Correction under various conditions are tabulated in Table 4.5 and Table 4.6.

4.3.1.1 Unbalanced Design (*J* = 4)

As per previous section, we will discuss performance of unbalanced design in the aspect of degree of heteroscedasticity, followed by the balanced design. Type I error results for unbalanced design when *J* = 4 are tabulated in Table 4.5.

4.3.1.1.1 Analysis on Degree of Heteroscedasticity

With regards to degree of heteroscedasticity, four groups of unequal sample sizes and group variances are used to obtain the Type I error rates as tabulated in Table 4.5.

Table 4.5

Type I error Rates for J = 4 Under Moderate and Extreme Degree of Heteroscedasticity

Distribution	**Variance**	**Pairing**	***P*-Mean**	**LSD with Bonfferoni correction**
Normal ***g* = 0, *h* = 0**	Moderate	Positive	0.05800	0.02800
		Negative	0.04150	0.20650
	Extreme	Positive	0.06950	0.02400
		Negative	0.05800	0.27850
Symmetric heavy tailed ***g* = 0, *h* = 0.5**	Moderate	Positive	0.05800	0.01250
		Negative	0.03650	0.16350
	Extreme	Positive	0.04450	0.01200
		Negative	**0.04850**	0.23500
Skewed normal tailed ***g* = 0.5, *h* = 0**	Moderate	Positive	0.06000	0.02900
		Negative	0.06850	0.44550
	Extreme	Positive	0.06050	0.03200
		Negative	0.07050	0.44600
Skewed heavy tailed ***g* = 0.5, *h* = 0.5**	Moderate	Positive	0.06000	0.04150
		Negative	**0.04750**	0.47150
	Extreme	Positive	0.06000	0.11850
		Negative	**0.05250**	0.51800
Grand Average			0.05588	0.19138

Indication:

▨	Type I error rates fulfill the Bradley's robust range
Bold	Type I error rates fulfill the robust stringent criterion

Grand average in the last row indicates the overall performance of *P*-Mean and LSD-Bonferroni correction across all types of distribution, from normal distribution to extreme distribution. From the grand average Type I error value, we notice that *P*-Mean has better overall performance as compared to LSD-Bonferroni correction. *P*-Mean generates Type I error rate (0.05588) which is only 0.00088 from the stringent upper limit of 0.05500. On the other hand, LSD-Bonferroni correction demonstrates failure in controlling Type I error rates with the grand average escalates up to 0.19138 value, far beyond the Bradley's upper limit. We will discuss on the performance of *P*-Mean and LSD-Bonferroni correction in detail in the following paragraphs.

Next, we move on to discuss on the performance of both procedures across all types of distribution. Firstly, under normal distribution, we observe that *P*-Mean generates Type I error rates which fell within the Bradley's robust interval regardless of the degree of heteroscedasticity and pairing types. Meanwhile, LSD-Bonferroni correction only manages to control Type I error rate under moderate degree of heteroscedasticity and positive pairing. It produces inflated Type I error rates of 0.20650 and 0.27850 under negative pairing for moderate and extreme degree of heteroscedasticity respectively. From the results, we can deduce that LSD-Bonferroni correction fails to control Type I error rate in the existence of heteroscedasity even though under normal distribution.

With regards to the performance of both procedures under symmetric heavy tailed distribution, again, *P*-Mean shows the capability to control Type I error rates effectively as compared to LSD-Bonferroni correction. Type I error rates generated by P-*M*ean for both moderate and extreme degree of heteroscedasticity met the Bradley's criterion of robustness with values in the range of 0.03650 to 0.05800. On the other hand, LSD-Bonferroni produces very conservative Type I error rate of 0.01250 and 0.01200 when combines positive pairing with moderate and extreme degree of heteroscedasticity respectively. Even under the combination of negative pairing with moderate and extreme degree of heteroscedasticity, LSD-Bonferroni still produce inflated rate of 0.16350 and 0.23500 respectively.

Under skewed normal tailed distribution, LSD-Bonferroni correction demonstrates its ability in controlling Type I error rates for positive pairing only under both moderate and extreme degree of heteroscedasticity, while generating inflated Type I error rate when the degree of heteroscedasticity is paired with negative pairing. In contrast, *P*-Mean remains effective in controlling Type I error rates across moderate and extreme degree of heteroscedasticity producing values in the range of 0.06000 to 0.07050. From here, again we can stress that only *P*-Mean shows ability in controlling Type I error rates successfully.

The last distribution in our study is skewed heavy tailed distribution. Here, *P*-Mean again demonstrates better performance as compared to LSD-Bonferroni correction for both moderate and extreme degree of heteroscedasticity. Under moderate degree of heteroscedasticity with positive pairing, Type I error rates for *P*-Mean and LSD-Bonferroni correction are 0.06000 and 0.04150 respectively. Both Type I error rates are within the Bradley's robust interval. For positive pairing under extreme degree of heteroscedasticity, *P*-Mean is robust with Type I error rate of 0.06000, while the rate for LSD-Bonferroni inflates to 0.11850. Meanwhile, under negative pairing of moderate and extreme degree of heteroscedasticity, the Type I error rates for *P*-Mean fulfill the stringent interval but not for LSD-Bonferroni correction, which produces inflated rates.

To conclude, *P*-Mean has shown the effectiveness in controlling the Type I error rates while LSD-Bonferroni correction failed in controlling Type I error.

4.3.1.2 Balanced Design (*J* = 4)

The results of *P*-Mean and LSD-Bonferroni correction for balanced design had been tabulated in Table 4.6. The results explained the robustness of both procedures under various conditions.

Table 4.6

Type I error Rates for J = 4 Balanced Design

Distribution	*P*-Mean	LSD with Bonfferoni correction
Normal ***g* = 0, *h* = 0**	0.06050	**0.04600**
Symmetric heavy tailed ***g* = 0, *h* = 0.5**	0.06350	0.02350
Skewed normal tailed ***g* = 0.5, *h* = 0**	0.05850	**0.04600**
Skewed heavy tailed ***g* = 0.5, *h* = 0.5**	0.05900	0.02300
Grand Average	0.06038	0.03463

Indication:

▨	Type I error rates fulfill the Bradley’s robust range
Bold	Type I error rates fulfill the robust stringent criterion

Grand Average value indicates that both *P*-Mean and LSD-Bonferroni correction are capable in controlling Type I error rates within the Bradley’s robust interval with the

Type I error value of 0.06038 and 0.03463 respectively. From Table 4.9, we notice that LSD-Bonferroni correction controlled Type I error rates more effective than *P*-Mean under normal distribution and skewed normal tailed distribution. LSD-Bonferroni correction produces Type I error rate of 0.04600 while *P*-Mean generates Type I error value of 0.06050 under normal distribution. For skewed normal tailed distribution, LSD-Bonferroni correction retains its Type I error at 0.04600 as in normal distribution, but the Type I error for *P*-Mean improved to 0.05850. For the remaining 2 types of distribution, which are symmetric heavy tailed distribution and skewed heavy tailed distribution, *P*-Mean demonstrates better performance of Type I error rates than LSD-Bonferroni correction. *P*-Mean generates Type I error rate of 0.06350 and 0.05900 under symmetric heavy tail distribution and skewed heavy tailed distribution respectively. In the meantime, LSD-Bonferroni correction produces the Type I error rate of 0.02350 and 0.02300 under symmetric heavy tailed distribution and skewed heavy tailed distribution respectively.

As expected, when all the assumptions are fulfilled, LSD-Bonferroni correction should perform at its best, and the result for normally distributed data with balanced design proved the notion. Nevertheless, under heavy tailed distribution, *P*-Mean has better control of Type I error rates.

Considering the performance of *P*-Mean and LSD-Bonferroni correction in unbalanced and balanced design for four group case, in general, we can conclude that, *P*-Mean surpasses LSD-Bonferroni correction in controlling Type I error rates.

4.3.2 Type I error for *J* = 6

The comparison between *P*-Mean and LSD-Bonferroni correction will be continued in the following sections for the six groups' case. All the results are tabulated in Table 4.7 and Table 4.8.

4.3.2.1 Unbalanced Design (*J* = 6)

Like in the previous section, for unbalanced design, the discussion will cover the performance of the two methods under the influence of heteroscedasticity and nature of pairings.

4.3.2.1.1 Analysis on Degree of Heteroscedasticity

Table 4.7 tabulates the Type I error rates of *P*-Mean and LSD-Bonferroni correction under moderate and extreme degree of heteroscedasticity for *J* = 6.

Table 4.7

Type I error Rates for J = 6 Under Moderate and Extreme Degree of Heteroscedasticity

Distribution	**Variance**	**Pairing**	***P*-Mean**	***LSD with Bonfferoni correction***
Normal ***g* = 0, *h* = 0**	Moderate	Positive	**0.04580**	0.03650
		Negative	0.06000	0.22100
	Extreme	Positive	**0.05000**	0.04350
		Negative	0.06000	0.31200
Symmetric heavy tailed ***g* = 0, *h* = 0.5**	Moderate	Positive	**0.05350**	0.02250
		Negative	0.05800	0.17650
	Extreme	Positive	**0.04800**	0.02700
		Negative	0.06300	0.27050
Skewed normal tailed ***g* = 0.5, *h* = 0**	Moderate	Positive	0.06650	0.03750
		Negative	**0.05350**	0.58550
	Extreme	Positive	**0.04900**	**0.05400**
		Negative	0.07350	0.55050
Skewed heavy tailed ***g* = 0.5, *h* = 0.5**	Moderate	Positive	0.06000	**0.05100**
		Negative	**0.05300**	0.61300
	Extreme	Positive	0.05700	0.01655
		Negative	0.06300	0.62700
Grand Average			0.05711	**0.22778**

Indication:

▨	Type I error rates fulfill the Bradley's robust range
Bold	Type I error rates fulfill the robust stringent criterion

The contrast in the Grand Average rate of Type I error for *P*-Mean (0.05711) and LSD-Bonferroni correction (0.22778) shows that the earlier manage to control Type I error rate more effectively than the latter. In general, we can assume that the *P*-Mean is more effective in controlling Type I error than the LSD-Bonferroni correction. A more detail analysis based on different distributions will be discussed next.

Under normal distribution, both *P*-Mean and LSD-Bonferroni correction generate Type I error rates which are robust for positive pairings under both moderate and extreme degree of heteroscedasticity. Nevertheless, when the pairing changes to negative, LSD-Bonferroni correction fails to control Type I error rates for both moderate and extreme degree of heteroscedasticity with inflated Type I error rates of 0.22100 and 0.31200 respectively.

When the distribution deviated from normal, such as in the case of symmetric heavy tailed, the pattern of the results is still the same. *P*-Mean performs better than LSD-Bonferroni correction as per Type I error rates depicted in Table 4.7. Again, *P*-Mean generates stringent Type I error rates under positive pairing for both variance type. On the other hand, LSD-Bonferroni correction only manage to return robust Type I error rate for positive pairing under extreme degree of heteroscedasticity. Under negative pairing, the Type I error rates for *P*-mean is 0.05800 for moderate and 0.06300 for extreme degree of heteroscedasticity. As expected, the rates for LSD-Bonferroni correction under negative pairing for both moderate and extreme degree of heteroscedasticity become liberal with values of 0.17650 and 0.27050 respectively.

P-Mean also performs better as compared to LSD-Bonferroni correction under skewed normal tailed distribution. However, LSD-Bonferroni correction manages to return stringent Type I error rate (0.05400) under positive pairing and extreme degree of heteroscedasticity. LSD-Bonferroni correction again produces liberal Type I error rates of 0.58550 and 0.55050 under negative pairing for moderate and extreme degree of heteroscedasticity respectively.

Last but not least, under skewed heavy tailed distribution, *P*-Mean still outperforms LSD-Bonferroni correction for both extreme and moderate degree of heteroscedasticity, except for positive pairing with moderate degree of heteroscedasticity where LSD-Bonferroni correction produces good Type I error of 0.05100.

Across all types of distributions and conditions investigated for unbalanced design with *J* = 6, we can conclude that *P*-Mean is always in control of Type I error whereas LSD-Bonferroni correction produces inflated Type I error rates in most cases. When we explore in details, we could observe that LSD-Bonferroni correction fail to control Type I error rate in all negative pairing and part of the positive pairing cases.

Overall, *P*-Mean generates more robust Type I error rates for unbalanced design as compared to LSD-Bonferroni correction when *J* = 6.

4.3.2.2 Balanced Design (*J* = 6)

Throughout this section, the performance of *P*-Mean and LSD-Bonferroni correction under balanced design will be discussed in detail.

Table 4.8

Type I error Rates for J = 6 Balanced Design

Distribution	***P*-Mean**	**LSD with Bonfferoni correction**
Normal ***g* = 0, *h* = 0**	0.07550	**0.04650**
Symmetric heavy tailed ***g* = 0, *h* = 0.5**	0.05700	0.03000
Skewed normal tailed ***g* = 0.5, *h* = 0**	0.07100	0.04250
Skewed heavy tailed ***g* = 0.5, *h* = 0.5**	0.05800	0.02650
Grand Average	0.06538	0.03638

Indication:

▨	Type I error rates fulfill the Bradley's robust range
Bold	Type I error rates fulfill the robust stringent criterion

The overall performance indicated by the Grand Average rates showed that both procedures successfully control Type I error rates within Bradley's interval with value of 0.06538 for *P*-Mean and 0.03638 for LSD-Bonferroni correction. With regards to distribution, from Table 4.8, we notice that LSD-Bonferroni correction (0.04650) performed better as compared to *P*-Mean (0.07550) under normal distribution. The Type I error rate generates by LSD-Bonferroni correction met the stringent robust interval while Type I error rate produced by *P*-Mean is slightly above the upper limit of Bradley's robust interval. For symmetric heavy tailed distribution, *P*-Mean generates Type I error rate of 0.05700 which is closer to the nominal value of 0.05 as compared to the Type I error rate generated by LSD-Bonferroni correction (0.03000). In contrast, under skewed normal tailed distribution, LSD-Bonferroni correction performs better than *P*-Mean with Type I error rate of 0.04250 and 0.07100 respectively. The Type I error rate for LSD-Bonferroni correction is closer to the nominal value. Meanwhile, under extreme condition, *P*-Mean (0.05800) had better control of Type I error rate than LSD-Bonferroni correction (0.02650).

Under the balanced design of *J* = 6, we can conclude that LSD-Bonferroni correction has better Type I error rates control under normal distribution and skewed normal tailed distribution while *P*-Mean controls Type I error rates more effectively under symmetric heavy tailed distribution and skewed heavy tailed distribution.

We observe that the pattern of the results generated for *J* = 4 and *J* = 6 balanced design are similar with LSD-Bonferroni correction controlled Type I error more effectively under normal distribution and skewed normal tailed distribution. On the other hand, *P*-Mean has the efficiency in Type I error controlling under symmetric heavy tailed distribution and skewed heavy tailed distribution.

In a nut shell, *P*-Mean has better control of Type I error rates as compared to LSD-Bonferroni correction in *J* = 6 case regardless whether the design is balanced or unbalanced

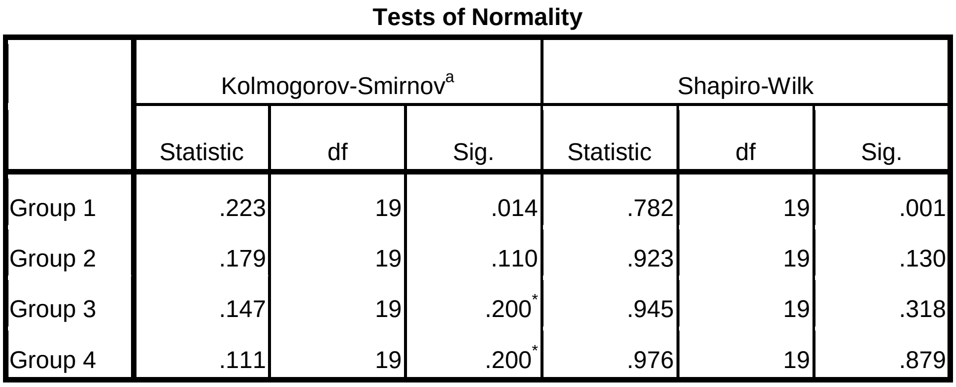
4.4 Real Data Analysis

To investigate on the effectiveness of the proposed methods on real data, we have collected marks scored by students in Decision Analysis courses which are conducted by four different lecturers. The effectiveness of the procedures proposed in testing the real group difference has been tested using the real data. We have also made the comparison of the new proposed procedures with the well-known classical LSD. The result has been tabulated in Table 4.11.

Before implementing the post hoc test, we carry out the normality test on each real data set by applying two common normality tests, namely Kolmogorov-Smirnov and Shapiro-Wilk tests. The significant values tabulate in Table 4.9 show that Group 2, Group 3 and Groups 4 are normally distributed while Group 1 is non-normally distributed.

Table 4.9

Normality Test



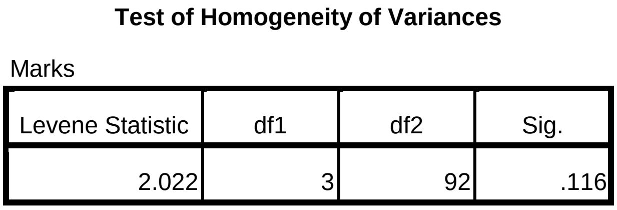
Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Group 1	.223	19	.014	.782	19	.001
Group 2	.179	19	.110	.923	19	.130
Group 3	.147	19	.200[*]	.945	19	.318
Group 4	.111	19	.200[*]	.976	19	.879

From Table 4.10, significant value returned by Levene Statistic is 0.116, which is greater than the significant level of 0.05. This denoted that the four groups have equal variance.

Table 4.10

Homogeneity of Variances Test



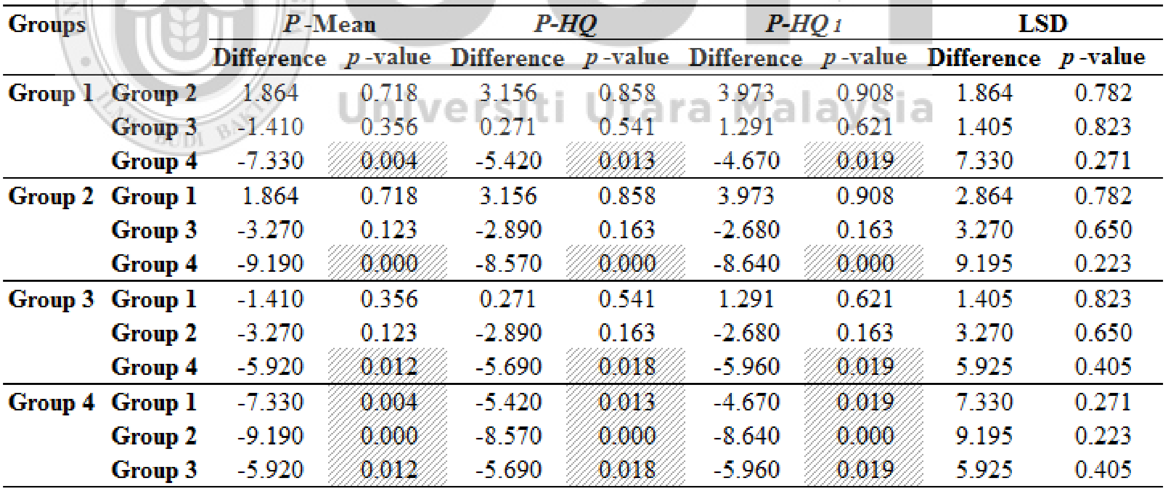
Test of Homogeneity of Variances

Marks

Levene Statistic	df1	df2	Sig.
2.022	3	92	.116

Table 4.11

p-value *Comparison for P-Mean, P-HQ, P-HQ$_1$ and LSD when J = 4*



Groups		***P*-Mean**		***P-HQ***		***P-HQ*$_1$**		**LSD**	
		Difference	***p*-value**	**Difference**	***p*-value**	**Difference**	***p*-value**	**Difference**	***p*-value**
Group 1	**Group 2**	1.864	0.718	3.156	0.858	3.973	0.908	1.864	0.782
	Group 3	-1.410	0.356	0.271	0.541	1.291	0.621	1.405	0.823
	Group 4	-7.330	0.004	-5.420	0.013	-4.670	0.019	7.330	0.271
Group 2	**Group 1**	1.864	0.718	3.156	0.858	3.973	0.908	2.864	0.782
	Group 3	-3.270	0.123	-2.890	0.163	-2.680	0.163	3.270	0.650
	Group 4	-9.190	0.000	-8.570	0.000	-8.640	0.000	9.195	0.223
Group 3	**Group 1**	-1.410	0.356	0.271	0.541	1.291	0.621	1.405	0.823
	Group 2	-3.270	0.123	-2.890	0.163	-2.680	0.163	3.270	0.650
	Group 4	-5.920	0.012	-5.690	0.018	-5.960	0.019	5.925	0.405
Group 4	**Group 1**	-7.330	0.004	-5.420	0.013	-4.670	0.019	7.330	0.271
	Group 2	-9.190	0.000	-8.570	0.000	-8.640	0.000	9.195	0.223
	Group 3	-5.920	0.012	-5.690	0.018	-5.960	0.019	5.925	0.405

Indication:

▨ *p*-value ≤ 0.05. True differences between groups detected

We may refer to Table 4.11 for post-hoc test result. The significant level has been set at 0.05. Null hypothesis is successfully being rejected if the *p*-value is less than or equal to the significant level. In other words, true differences between groups are said to exist if the *p*-value is less than or equal to the significant level.

From Table 4.11, all *P*-Mean, *P-HQ* and *P-HQ*$_1$ successfully identify 3 significant group's difference between group 4 and the other groups. *P*-Mean produced the smallest *p*-value, followed by *P-HQ* and *P-HQ*$_1$. However, LSD fails to detect any of the groups' difference.

We can conclude that compared to the classical procedure LSD, all *P*-Mean, *P-HQ* and *P-HQ*$_1$ have shown that they are capable to detect the true difference between groups effectively.

CHAPTER FIVE
CONCLUSION

5.1 Introduction

The goal of this study is to develop the pairwise MCPs for hinge estimators by using the *P*-Method proposed by Wilcox (2001, 2003). Wilcox (2001) has proved that *P*-Method has good control of Type I error rates when tested on modified one step M-estimator (MOM). It could ensure that the Type I error rates does not exceed the nominal value as well as does not drop too far from the nominal value. Out of the seven hinge estimators proposed by Reed and Stark (1996), *HQ* and HQ_1 are recommended as the most robust estimators (Keselman, Wilcox, Lix, Algina & Fradette, 2007). Thus, these two estimators are chosen for further development for post hoc analysis. *HQ* and HQ_1 are adaptive asymmetric trimmed means. The trimming is based on the characteristic of a distribution's tail. These two estimators are proven to have good control in Type I error rates when used as central tendency measures in Alexander Govern test (Abdullah, Syed Yahaya & Othman, 2010).

All the investigated estimators namely, the Mean, HQ and HQ_1 are then paired with *P*-Method and tested for the effect of non-normality and heteroscedasticity on the Type I error. Balanced and unbalanced conditions have been designed for different number of groups (varying from two to four groups) in order to test the effectiveness of *P*-Mean, *P-HQ* and *P*-HQ_1 procedures. Four types of distributions such as normal distribution, symmetric heavy tailed distribution, skewed normal tailed distribution and skewed heavy tailed distribution are used in this study with the purpose to test the effect of the non-normality towards the Type I error. Besides, unequal variances of 1:36 are also used to test for the heteroscedasticity effect. Other variables such as nature of pairings of group sizes and group variances (positive and negative pairing) are also taken into the count as these variables were proven to have some effect on the Type I error rates for unbalanced design condition.

The simulated data sets are generated using *SAS/IML* Version 9.2. Regardless of the specifications, each procedure is simulated 2,000 times and then bootstrapped for 2,000 times as well. The bootstrap percentile method is used to test the hypothesis due to the intractability of the sampling distributions of the statistics.

The Type I error rates are determined and compared. We compared the Types I Error rates between *P*-Mean, *P-HQ* and *P-HQ*$_1$ procedures. The best procedure is the one which could produce the Type I error rates closest to the nominal value of 0.05. Bradley's liberal criterion of robustness is used to benchmark the robustness of the procedures. A procedure is considered as robust if its empirical rates of Type I error is within the interval of $0.025 \leq ά \leq 0.075$ under a 0.05 significant level.

5.2 *P*-Method Performance

The performance of *P*-Method with different type of estimators will be discussed in Section 5.2.1 and Section 5.2.2 below. From this section, we can figure out the best procedure under each distribution.

5.2.1 Conditions for Distributions

In total, 120 conditions are being designed for both J = 4 and J = 6, with 10 conditions for each estimators and each type of distributions. For instance, there are

10 conditions designed for *P*-Mean under normal distribution. Performance of *P*-Method when associated with each estimator is summarized in Table 5.1 in the form of percentage with frequency in the bracket.

Table 5.1

Percentage of Robust Conditions for P-Method under J = 4 and J = 6

Distribution	**Procedures**		
	***P*-Method**		
	Mean	***HQ***	***HQ_1***
Normal ***g* = 0, *h* = 0**	90% (9/10)	80% (8/10)	100% (10/10)
Symmetric heavy tailed ***g* = 0, *h* = 0.5**	100% (10/10)	100% (10/10)	80% (8/10)
Skewed normal tailed ***g* = 0.5, *h* = 0**	100% (10/10)	100% (10/10)	100% (10/10)
Skewed heavy tailed ***g* = 0.5, *h* = 0.5**	100% (10/10)	100% (10/10)	100% (10/10)

Indication:

▨ 100% robustness in the conditions investigated

As could be observed in Table 5.1, 90% of the conditions investigated are robust for Mean. From all the conditions, 1 out of 10 conditions is not robust when Mean is put to test under normal condition in the case of *J* = 6. *HQ* has 80% of the conditions to be robust, and is not robust in both positive and negative pairing with extreme degree of heterogeneity when J = 6. From Table 5.1, we noticed that HQ_1 is the best performer under normal distribution with the achievement of 100% robustness.

Next, for symmetric heavy tailed distribution, both Mean and HQ achieved 100% robustness for J = 4 and J = 6. Percentage of robust conditions for HQ_1 is only 80%. The non-robustness occurs in positive pairing for both moderate and extreme degree of heterogeneity when J = 6.

Under skewed normal tailed distribution, all Mean, HQ and HQ_1 perform perfectly by producing 100% robust conditions.

For skewed heavy tailed distribution, all Mean, HQ and HQ_1 again show their effectiveness in Type I error controlling by achieving 100% robustness.

From Table 5.1, we can conclude that,

i. HQ_1 achieves 100% robustness for all the conditions investigated in normal distribution

ii. Mean and HQ achieves 100% robustness for all the conditions investigated in symmetric heavy tailed distribution

iii. Mean, HQ and HQ_1 achieves 100% robustness for all the conditions investigated in skewed normal tailed distribution

iv. Mean, HQ and HQ_1 achieves 100% robustness for all the conditions investigated in skewed heavy tailed distribution

5.2.1.1 Four Groups Case (*J* = 4)

Table 5.2 summarizes the percentage of robust conditions of *P*-Method for *J* = 4 under balanced and unbalanced design. There are 60 conditions investigated under this case, with 15 conditions for each type of distribution.

Table 5.2

Percentage of Robust Conditions for P-Method under J = 4

Distribution	**Procedures**		
	***P*-Method**		
	Mean	***HQ***	***HQ_1***
Normal ***g* = 0, *h* = 0**	100% (5/5)	100% (5/5)	100% (5/5)
Symmetric heavy tailed ***g* = 0, *h* = 0.5**	100% (5/5)	100% (5/5)	100% (5/5)
Skewed normal tailed ***g* = 0.5, *h* = 0**	100% (5/5)	100% (5/5)	100% (5/5)
Skewed heavy tailed ***g* = 0.5, *h* = 0.5**	100% (5/5)	100% (5/5)	100% (5/5)

Indication:

▨ 100% robustness in the conditions investigated

From Table 5.2, we can conclude that for the case of *J* = 4 all the three estimators meet the Bradley’s criterion of robustness regardless of the conditions tested.

5.2.1.2 Six Groups Case (*J* = 6)

The robustness of *P*-Method for both balanced and unbalanced under *J* = 6 is recapped in Table 5.3. Same conditions as *J* = 4 have been tested.

Table 5.3

Percentage of Robust Conditions for P-Method under J = 6

Distribution	**Procedures**		
	***P*-Method**		
	Mean	***HQ***	***HQ_1***
Normal ***g* = 0, *h* = 0**	80% (4/5)	60% (3/5)	100% (5/5)
Symmetric heavy tailed ***g* = 0, *h* = 0.5**	100% (5/5)	100% (5/5)	60% (3/5)
Skewed normal tailed ***g* = 0.5, *h* = 0**	100% (5/5)	100% (5/5)	100% (5/5)
Skewed heavy tailed ***g* = 0.5, *h* = 0.5**	100% (5/5)	100% (5/5)	100% (5/5)

Indication:

▨ 100% robustness in the conditions investigated

Under normal distribution, the percentage of non-robust conditions for Mean takes place under balanced design while for *HQ*, it happens under both pairings with extreme degree of heterogeneity. HQ_1 is the only estimator produces 100% robust condition.

For the symmetric heavy tailed distribution, only *HQ_1* shows some percentage of non-robustness, and it is traced under positive pairing with moderate and extreme degree of heterogeneity.

For skewed normal tailed and skewed heavy tailed distribution, the three estimators Mean, *HQ* and *HQ_1* studied in our research show the capability to control Type I error rates by producing 100% robustness.

From Table 5.3, we can conclude that,

i. *HQ_1* achieves 100% robustness for all the conditions investigated in normal distribution

ii. Mean and *HQ* achieve 100% robustness for all the conditions investigated in symmetric heavy tailed distribution

iii. Mean, *HQ* and *HQ_1* achieve 100% robustness for all the conditions investigated in skewed normal tailed distribution

iv. Mean, *HQ* and *HQ_1* achieve 100% robustness for all the conditions investigated in skewed heavy tailed distribution.

5.2.2 Conditions for Heteroscedasticity

In this section, we tabulate the robustness result accordingly to the study design. As mentioned earlier, there are 120 conditions being investigated in this study. For further breakdown, there are 96 conditions created for unbalanced design while 24 conditions created for balanced design. For each procedure, there are 32 conditions under unbalanced design while 8 conditions are under balanced design. Balanced design consists of conditions with equal variances and equal sample sizes.

Meanwhile, the conditions for unbalanced design are combinations of unequal sample sizes with unequal variances.

Table 5.4

P-Method Robustness in Balanced and Unbalanced Design

Distribution	**Procedures**		
	***P*-Method**		
	Mean	***HQ***	***HQ_1***
Balanced	87.5% (7/8)	▨ 100% (8/8)	▨ 100% (8/8)
Unbalanced	▨ 100% (32/32)	93.8% (30/32)	93.8% (30/32)

Indication:

▨ 100% robustness in the conditions investigated

We notice that the performance of *HQ* and *HQ_1* are best for balanced design with all the conditions meet the Bradley's robustness criterion. On the other hand, for unbalanced design, Mean shows the highest percentage (100%) of robust conditions while *HQ* and *HQ_1* produce 93.8% respectively.

In general, all estimators are proven to have good control of Type I error rates across all the designs except for application of *P-HQ* and *P-HQ_1* under unbalanced design of J = 6. The outcome of this study is being summarized in Figure 5.1.

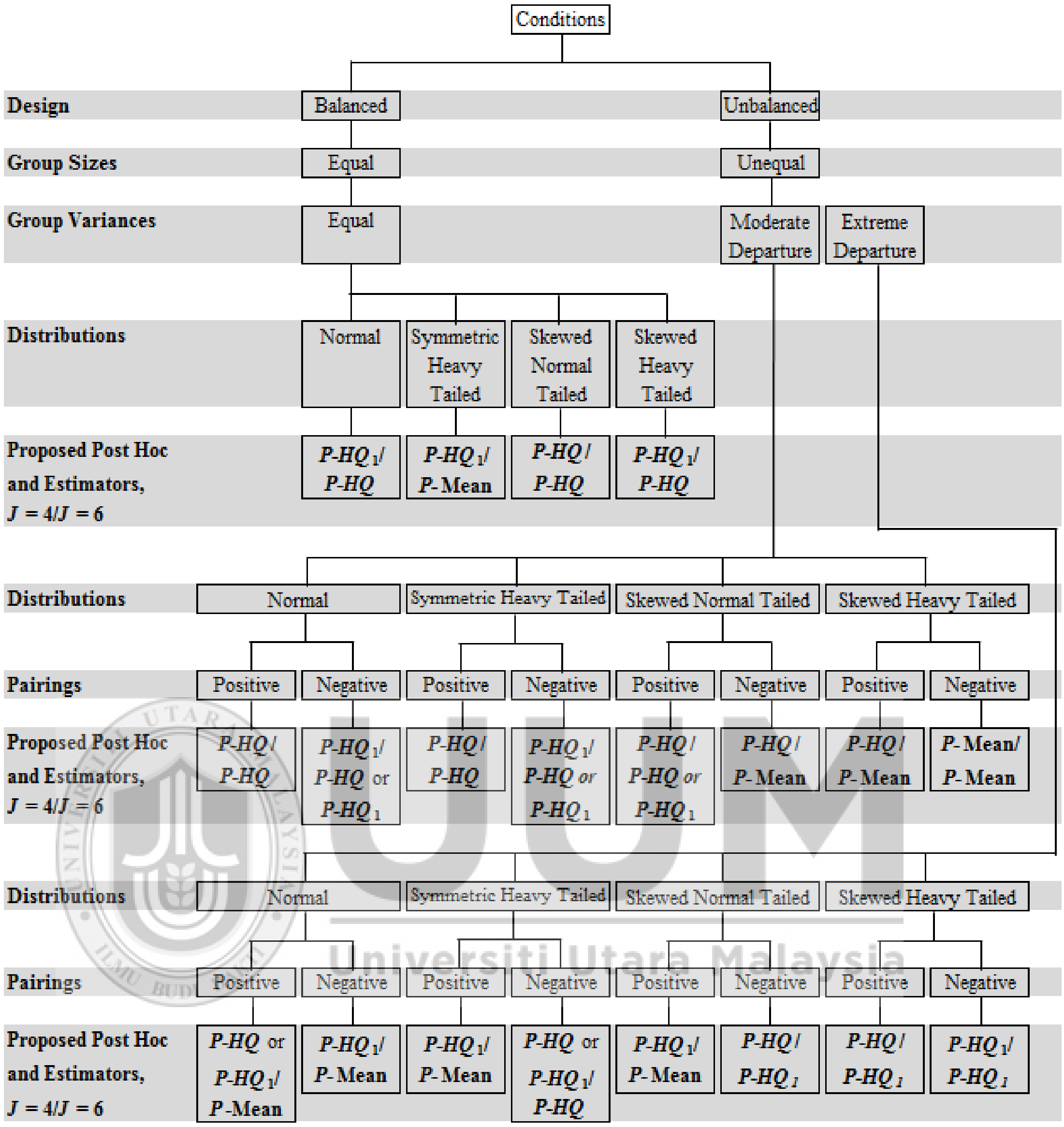


Figure 5.1. Summary of proposed post hoc test and estimators

5.3 *P*-Mean and LSD-Bonferroni Correction in a Nut Shell

Section 5.3.1 and Section 5.3.2 below indicate the comparison of performance between *P*-Mean and LSD with Bonferroni correction.

5.3.1 Conditions for Distributions

For *P*-Mean and LSD-Bonferroni correction, a total of 80 conditions have been designed for both *J* = 4 and *J* = 6, with 10 conditions for each procedures under each type of distribution. Table 5.5 summarizes the performance of *P*-Mean and LSD-Bonferroni correction in the form of percentage and robust frequency in the bracket.

Table 5.5

Percentage of Robust Conditions for P-Mean and LSD-Bonferroni Correction under J = 4 and J = 6

Distribution	***P*-Mean**	**LSD with Bonfferoni correction**
Normal ***g* = 0, *h* = 0**	90% (9/10)	50% (5/10)
Symmetric heavy tailed ***g* = 0, *h* = 0.5**	100% (10/10)	20% (2/10)
Skewed normal tailed ***g* = 0.5, *h* = 0**	100% (10/10)	60% (6/10)
Skewed heavy tailed ***g* = 0.5, *h* = 0.5**	100% (10/10)	30% (3/10)

Indication:

▨ 100% robustness in the conditions investigated

From Table 5.5, when under normal distribution, both *P*-Mean and LSD-Bonferroni correction fail to achieve 100% robust condition. However, *P*-Mean (90%) manages to achieve higher robust condition percentage as compared to LSD-Bonferroni correction (50%), with only 1 failure in controlling the Type I error rate.

Next, for symmetric heavy tailed distribution, *P*-Mean achieves 100% robustness while LSD-Bonferroni correction only manages to achieve 20% of the robust conditions.

P-Mean again achieves 100% robust conditions under skewed normal tailed distribution. On the other hand, LSD-Bonferroni correction robust conditions' is at 60%, which is the best achievement as compared to other distributions achievements.

For skewed heavy tailed distribution, similar to percentage of robust conditions achieved by *P*-Mean and LSD-Bonferroni correction under symmetric heavy tailed, *P*-Mean achieves 100% robustness while LSD-Bonferroni correction only manages to attain 30% of the robust conditions.

From Table 5.5, we can summarize that, *P*-Mean achieves 100% robustness for all the conditions in all types of distribution except for normal distribution.

5.3.1.1 Four Groups Case (J = 4)

The percentage of robust conditions for *P*-Mean and LSD-Bonferroni correction when J = 4 are tabulated in Table 5.6. There are 40 conditions investigated in this case, with 5 conditions for each type of procedure and distribution.

Table 5.6

Percentage of Robust Conditions for P-Mean and LSD-Bonferroni Correction under J = 4

Distribution	***P*-Mean**	**LSD with Bonfferoni correction**
Normal ***g* = 0, *h* = 0**	100% (5/5)	40% (2/5)
Symmetric heavy tailed ***g* = 0, *h* = 0.5**	100% (5/5)	0% (0/5)
Skewed normal tailed ***g* = 0.5, *h* = 0**	100% (5/5)	60% (3/5)
Skewed heavy tailed ***g* = 0.5, *h* = 0.5**	100% (5/5)	20% (1/5)

Indication:

▨ 100% robustness in the conditions investigated

P-Mean achieves 100% robustness for the case of *J* = 4 regardless of the type of distributions tested (refer to Table 5.6). On the other hand, LSD-Bonferroni correction achieves highest percentage of robust conditions under skewed normal tailed distribution (60%), followed by normal distribution (40%) and lastly by skewed heavy tailed distribution (20%). Meanwhile, LSD-Bonferroni correction totally fails under symmetric heavy tailed distribution.

Thus, the study shows that *P*-Mean outperforms LSD-Bonferroni correction in all type of distributions when *J* = 4.

5.3.1.2 Six Groups Case (*J* = 6)

The overall performance of *P*-Mean and LSD-Bonferroni correction in the case of *J* = 6 is summarized in Table 5.7. Similar to *J* = 4 case, a total 40 conditions are investigated, with 5 conditions for each type of procedure and distribution.

Table 5.7

Percentage of Robust Conditions for P-Mean and LSD-Bonferroni Correction under J = 6

Distribution	***P*-Mean**	**LSD with Bonfferoni correction**
Normal ***g* = 0, *h* = 0**	80% (4/5)	60% (3/5)
Symmetric heavy tailed ***g* = 0, *h* = 0.5**	100% (5/5)	40% (2/5)
Skewed normal tailed ***g* = 0.5, *h* = 0**	100% (5/5)	60% (3/5)
Skewed heavy tailed ***g* = 0.5, *h* = 0.5**	100% (5/5)	40% (2/5)

Indication:

▨ 100% robustness in the conditions investigated

Table 5.7 shows that *P*-Mean achieves 100% robustness under all types of distribution except for normal distribution. For normal tailed distributions, *P*-Mean only manages to achieve 80% robust conditions. LSD-Bonferroni correction has the equal performance under normal distribution (60%) and skewed normal tailed distribution. The percentage of robust conditions for symmetric heavy tailed and skewed heavy tailed distributions are the same with each having 40% robust conditions.

To put it in a nut shell, *P*-Mean is still the best choice in all type of distributions as compared to LSD-Bonferroni correction.

5.3.2 Conditions for Heteroscedasticity

Tabulated in Table 5.8 are the overall performance of *P*-Mean and LSD-Bonferroni correction according design of the study. As mentioned earlier, a total of 80 conditions being investigated for *P*-Mean and LSD-Bonferroni correction with 64 conditions for unbalanced design and 16 conditions for balanced design. For each procedure, there are 32 conditions under unbalanced design while 8 conditions created under balanced design.

Table 5.8

P-Mean and LSD-Bonferroni Correction Robustness in Balanced and Unbalanced Design

Distribution	***P*-Mean**	**LSD with Bonfferoni correction**
Balanced	87.5% (7/8)	75% (6/8)
Unbalanced	100% (32/32)	31.3% (10/32)

In both balanced and unbalanced design, *P*-Mean achieves the highest robust condition with the percentage of 87.5% and 100% respectively while LSD-Bonferroni correction achieves 75% robust conditions under balanced design and 31.3% for unbalanced design. Design of the study shows some impact on LSD-

Bonferroni correction whereby this procedure performs better under balanced design as compared to unbalanced design. In contrast, the small disparity between the designs for *P*-Mean indicates that this procedure is robust to the study design.

In general, we can conclude that *P*-Mean outperforms LSD-Bonferroni correction in all types of conditions, regardless of cases.

5.4 *P*-Mean Robustness

One of our significant findings in this study is the outstanding performance in controlling Type I error rates when using *P*-Method to test for means (*P*-Means). At the early stage of this study, we would expect that the performance of *HQ* and HQ_1 using *P*-Method will bring more significant robustness as compared to *P*-Mean. However, this research has proven that *P*-Mean remains robust in both balanced and unbalanced design, regardless of the type of distributions for both J = 4 and J = 6.

5.5 Overall Summary

In this section, we will summarize the performance of all the procedures, *P*-Mean, *P-HQ*, *P*-HQ_1 and LSD-Bonferroni correction. In total, there are 160 conditions being investigated in this study with 40 conditions each designed for *P*-Mean, *P-HQ*, *P*-HQ_1 and LSD-Bonferroni correction. In both J = 4 and J = 6 cases, the best procedure to control Type I error rates is *P*-Mean with the robust condition of 97.5% (39/40) and the first runner up is *P-HQ* with 2.5% left behind *P*-Mean. The third best procedure is *P*-HQ_1 with the robust condition of 95%, followed by LSD-Bonferroni correction with the robust condition of 40%.

LSD is chosen for this research because it is among the earliest post-hoc procedure to be introduced and it is also simple to calculate. Furthermore, due to its reliability, this procedure is available in most of the software packages. To improve the performance of this procedure, in this study we apply Bonferroni correction. However, the improvement is still far below satisfaction, and it appears to be the worst procedure in this study.

In the initial stage of this study, we expect *P*-Mean to be among the worst procedures as the classical mean is known to be sensitive under non normality, but ironically, our finding shows that this procedure turns out to be the most effective estimator in controlling Type I error even under severe conditions. This implies that the classical mean is able to control Type I error rates effectively when uses with *P*-Method and it remains its robustness across all types of distribution. It even outperformes the robust estimators in this study. This finding is beneficial to the users of statistics, especially to those who tend to skip the assumptions checking. However, since post hoc test is a follow up of omnibus test, so, a further study on the omnibus test with regards to means should be done. Even though *P-HQ* and *P-HQ*$_1$ are not the best performer in this study, but they produce reasonable Type I error rates, thus, we can suggest that these post hoc procedures to be applied together with omnibus tests using any of the corresponding estimators, *HQ* or *HQ*$_1$. Among the omnibus tests available to be used with these procedures are Alexander Govern with *HQ* and *HQ*$_1$ (Abdullah, 2011) and *H* test with *HQ* and *HQ*$_1$ (Muhammad Di, 2013).

5.6 Implications

Our goal in this study is to develop the program for robust Multiple Pairwise Comparison Procedure for *HQ* and HQ_1. Besides, we also evaluate and compare the performance of the *P-HQ* and $P\text{-}HQ_1$ procedures with *P*-Mean and LSD-Bonferroni correction procedures. In this section, we would like to share some of the findings that emerged from this study.

Generally, we found that when under unbalanced design for both $J = 4$ and $J = 6$ cases, *P-HQ* performs as well as $P\text{-}HQ_1$ with robust conditions of 93.8% respectively. Our analysis also found that *P*-Mean performs remarkably well under unbalanced design for both $J = 4$ and $J = 6$ cases across all types of distribution. Out of the 32 conditions studied, *P*-Mean achieves 100% robust condition. On the other hand, LSD-Bonferroni correction shows the worst performance with robust condition of 31.3%.

For balanced design, both *P-HQ* and $P\text{-}HQ_1$ show the equal capability in controlling Type I error rates. Both procedures score 100% robust conditions for all the 8 conditions investigated respectively. Under balanced design, *P*-Mean fails to compete with *P-HQ* and $P\text{-}HQ_1$ with only 87.5% robust conditions. The failure is spotted in normal distribution. For LSD-Bonferroni correction, even though this procedure still rank the last among all the procedures, it has better control of Type I error rates as compared to unbalanced design performance with 75% robust conditions.

In general, after considering the performance of both *P-HQ* and *P-HQ*$_1$ under various conditions, we can deduce that both *P-HQ* and *P-HQ*$_1$ have equal effectiveness in controlling Type I error. This is due to the fact that *P-HQ* and *P-HQ*$_1$ produce the same number of robust conditions for both balanced and unbalanced design. This study also revealed that the classical mean works well using *P*-Method. This could be an alternative procedure to alleviate the problem of non-robustness when working with classical mean.

5.7 Suggestions for Future Research

In this research, our interest is on *P-HQ* and *P-HQ*$_1$ procedures in terms of robustness. We have proved that *HQ* and *HQ*$_1$ estimators when used with *P*-Method are able to generate good results across most of the distributions. However, there are still rooms for improvement especially under unbalanced design. There are a few other multiple comparison procedures as suggested by Wilcox (2001) such as *PW*-Method and *PTW*-Method which could improve the control of Type I error rates for *HQ* and *HQ*$_1$ across all types of distributions.

Since *P*-Method is proven to perform well even on classical mean, the researchers can also consider adopting other robust location estimators on this method for post hoc purposes. A few examples are the hinge estimators proposed by Reed and Stark (1996) such as *HQ*$_2$, *HH*$_1$, *HH*$_3$, *HSK*$_2$ and *HSK*$_5$.

The findings also reveals that by using *P*-Method, Type I error rates generated for unbalanced design are better when compared to balanced design when the classical mean estimator is used. This situation occurred on normal distribution in $J = 6$ cases. When *HQ* and HQ_1 estimators are applied on *P*-Method, they fail to achieve 100% robust conditions under unbalanced design. The failure is detected under normal distribution and skewed heavy tailed distribution for *HQ* and HQ_1 respectively.

The aforementioned limitations should not be neglected if our goal is to search for a multiple pairwise comparison procedure which can effectively control Type I error rates (robust) as well as perform consistently in both balanced and unbalanced design.

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* (8thed.). 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