NEW POWER ANALYSIS FOR THE PSEUDO-MEDIAN PROCEDURE FOR MORE THAN TWO GROUPS

by

Ibtesam Ali Alsaggaff

Thesis submitted in fulfillment of the

requirements for the degree

of Doctor of Philosophy

June 2013

ACKNOWLEDGEMENTS

This thesis would not have been possible without the guidance and the help of several individuals who in one way or another contributed and extended their valuable assistance in the preparation and completion of this study.

My great gratitude to my main supervisor, Professor Low Heng Chin, for her sincerity, patience and steadfast encouragement to complete this study. Her constructive comments and invaluable suggestions have contributed to the success of this research.

My deepest appreciation to my co-supervisor, Professor Abdul Rahman Othman, for his supervision and knowledge regarding this topic. I can't say thank you enough for his tremendous support and help. Without his encouragement and guidance this thesis would not have materialized.

Gratitude to Professor Padmanabhan, Visiting Professor from Monash University, Australia for his suggestion and initial guidances for this research. Gratitude and appreciation to Jessica, Aishah, Sin Yin and also to all those who have contributed to this research, directly or indirectly, or gave me moral support during my study

Deep gratitude and appreciation to my husband for his love, patience and standing beside me in my weakness and my strength throughout my study and my life. My sincere thanks to my parents for their love, prayers and constant encouragement to complete my studies. Finally, I do not forget my wonderful children, brothers, sisters and my family members for their continuous prayer and encouragement throughout my study.

TABLE OF CONTENTS

CHAPTER 4 - RESULTS OF THE ANALYSIS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

APPENDIX A

APPENDIX B

APPENDIX C

LIST OF PUBLICATIONS

1. Alsaggaff, I. A., Othman, A. R., & Low, H. C. (2012). Estimated power of a robust method for treatment groups comparison. In *American Institute of Physics (AIP) Conference Proceedings*, Volume **1482**, pp. 502-506.

2. Alsaggaff, I. A., Othman, A. R. & Low, H. C. (2012). *Type I Error of the Pseudo Median Procedure in the One Way ANOVA Unbalanced Design.* Paper presented at the the 2nd Regional Conference on Applied and Engineering Mathematics (RCAEM 2012), 30-31 May, Penang, Malaysia.

3. Alsaggaff, I. A., Othman, A. R. & Low, H. C. (2011). *Type I Error and Power Curve of the Pseudo Median Procedure For Unbalanced Design.* In Proceeding of the 10th International Annual Symposium (UMTAS 2011), 11-13 July, Kuala Terengganu, Malaysia, pp. 27-33

4. Alsaggaff, I. A., Othman, A. R. & Low, H. C. (2010). *Detecting Differences in more than Two Groups using Pseudo-Medians.* In Proceedings of the 2nd International Conference on Mathematical Sciences (ICMS2 2010) [CD ROM], 30 Nov - 3 Dec, Kuala Lumpur, Malaysia.

5. Alsaggaff, I. A., Othman, A. R. & Low, H. C. (2012). *Type I Error and New Power Analysis for Pseudo-Median Procedure.* Submitted to the Communications in Statistics - Simulation and Computation.

ANALISIS KUASA BAHARU BAGI PROSEDUR PSEUDO-MEDIAN UNTUK LEBIH DARIPADA DUA KUMPULAN

ABSTRAK

Perbandingan kumpulan rawatan adalah kerap digunakan dalam penyelidikan praktikal dalam pelbagai bidang. Ujian berparameter ANOVA *F* adalah yang paling meluas digunakan untuk membandingkan kumpulan rawatan, khususnya min bagi tiga atau lebih kumpulan rawatan. Walau bagaimanapun, ujian berparameter biasanya memerlukan andaian kenormalan dan kehomogenan varians. Jadi, kegagalan dalam andaian-andaian tersebut membawa kepada herotan ralat Jenis I dan pengurangan yang ketara dalam kuasa ujian. Oleh itu, prosedur pseudo-median yang menggunakan pseudo-median sebagai parameter lokasi telah dibangunkan untuk perbandingan kumpulan rawatan. Prosedur ini adalah pengubahsuaian prosedur tak berparameter Wilcoxon satu sampel yang dibangunkan untuk lebih daripada dua kumpulan. Prosedur pseudo-median adalah penjumlahan perbandingan berpasangan berganda antara kumpulan kawalan dan setiap kumpulan rawatan. Dalam kajian ini, prestasi prosedur pseudo-median diukur apabila andaian kenormalan dan keheterogenan tidak dipatuhi. Ralat Jenis I diperiksa dan satu analisis baru bagi kuasa prosedur ini dicadangkan dan dijalankan untuk lebih daripada dua kumpulan. Kedua-dua ralat Jenis I dan analisis kuasa dilakukan di bawah pelbagai darjah kehomogenan dan bentuk taburan yang berbeza. Kaedah butstrap digunakan untuk menjana taburan pensampelan pseudo bagi statistik ujian pseudo-median. Prestasi prosedur ini juga dibandingkan dengan ujian-ujian klasik iaitu ujian *F* ANOVA dan ujian pangkat hasiltambah Kruskal-Wallis.

Prosedur pseudo-median menunjukkan beberapa kekuatan dalam prestasinya terutama dalam mengawal ralat Jenis I. Ia mendemonstrasikan keteguhan untuk mengawal ralat Jenis I dengan situasi yang berbeza daripada ketidaknormalan, keheterogenan dan juga apabila saiz sampel adalah tidak sama dan berpasangan secara negatif dengan varians kumpulan. Kuasa prosedur ini tidak dipengaruhi oleh jenis taburan. Walau bagaimanapun, ia dipengaruhi oleh varians yang berbeza. Prosedur ini boleh memberi kuasa yang tinggi selagi varians adalah sama. Kuasa yang lemah dicerap apabila bilangan perbandingan berpasangan antara setiap kumpulan rawatan dan kawalan dengan saiz kesan bukan sifar adalah kecil. Kuasa bagi ujian-ujian tradisional dipengaruhi oleh jenis taburan. Kuasa adalah lemah apabila taburan adalah terpencong dengan ekor panjang, namun pseudo-median memberikan kuasa yang lebih tinggi untuk taburan jenis ini.

NEW POWER ANALYSIS FOR THE PSEUDO-MEDIAN PROCEDURE FOR MORE THAN TWO GROUPS

ABSTRACT

Comparison of treatment groups is frequently used in practical research in a variety of fields. The parametric ANOVA *F* test is most widely used to compare groups of treatment, specifically the means of three or more treatment groups. However, the parametric test usually requires normality of the distribution and homogeneity of variances. So, failure in meeting these assumptions leads to distortion of Type I error and substantial reduction in the power of the test. Therefore, the pseudo-median procedure which adopts the pseudo-median as a location parameter was developed for treatment groups comparison. This procedure is a modification of the one-sample nonparametric Wilcoxon procedure developed for more than two groups. The pseudo-median procedure is a summation of multiple paired comparisons between the control group and each of the treatment groups. In this study, the performance of the pseudo-median procedure is examined when the assumptions of normality and heterogeneity are violated. The Type I error is examined and a new power analysis of this procedure is proposed and carried out for more than two groups. Both Type I error and power analysis are performed under various degrees of homogeneity and different shapes of distributions. The bootstrap method is employed to generate a pseudo sampling distribution for the pseudo-median test statistic. The performance of this procedure is also compared against those of the classical tests, i.e., ANOVA *F* test and Kruskal-Wallis sum rank test. The pseudo-median procedure shows a number of strengths in its

performance especially in controlling the Type I error. It demonstrates its robustness to control the Type I error under different situations from non-normality, heterogeneity as well as when sample sizes are unequal and are negatively paired with group variances. The power of this procedure is not affected by the shape of the distribution. However, it is affected somewhat by the heterogeneity of variances. This procedure can provide high power so long as the variances are equal. Poor power was observed when the number of pairwise comparisons between each treatment group and control with non-zero effect size is small. The power of traditional tests is affected by the type of the distribution. The power is poor when the distribution is skewed with long tail, yet the pseudo-median provides much higher power for this type of the distribution.

CHAPTER 1

INTRODUCTION

1.1 Background of the study

The comparison of quantitative characteristics of two or more groups is commonly found in most statistical applications. The quantitative characteristic usually used in this comparison is the central tendency measure, specifically the mean. The mean is the location parameter compared in the parametric tests such as Student *t*-test and the ANOVA *F*-test. These parametric tests usually have high power. However, achieving this requires verification of certain data assumptions imposed by the tests. Unfortunately, these assumptions usually fail with real data.

The normality of data and homogeneity of variances are the usual assumptions for these parametric tests. The violation of these assumptions harms the performance of the parametric tests. Departure from the assumptions substantially inflates the Type I error and reduces the power.

Robust statistics came into being to deal with the problem of deviations from the assumptions. Robust statistics provide alternative procedures insensitive against the violation of the assumptions. The theory of robust statistics started more than 40 years ago (Ronchetti, 2006). Huber (1964) and Hampel (1968) provided the fundamental concept of robust statistics and gave the foundation of modern robust statistics. Huber developed the first robust estimator for the location parameter and Hampel followed that by deriving the influence function of an estimator. This was considered as an important

characteristic for a robust estimator. Robust statistics is still an active area of research. Ronchetti (2006) found 1617 papers on robust statistics in statistical journals. Many more can be found in journals of other fields where robust statistics was applied.

The statistical tests based on robust estimators are alternatives to the parametric tests, developed to provide better control of Type I error and high power when the assumptions are invalid. The researchers need to examine the performance of the proposed and developed tests in terms of Type I error and power. This adds to the knowledge on the validity of using these developed tests under different conditions of heterogeneity and non-normality.

With regards to the evaluation of the power, researchers usually estimate power at a few points which do not reflect the actual performance of a test (Babu & Padmanabhan, 2002; Babu *et al.*, 1999). In addition, the manner of the test statistic for comparing groups of treatment is different from test to test. Some tests compare groups as a collection of multiple pairwise comparisons between two groups while other tests compare each group with the combined groups. Therefore, the power analysis should be different depending on the manner of the test statistic for comparing treatment groups in order to give proper evaluation for the performance of the test. However, researchers still use the same power analysis for all treatment groups comparison (Keselman *et al.*, 2004b; Othman *et al.*, 2004a).

In traditional tests, power was examined by imposing the differences between each group and combined groups. In ANOVA *F* test, the comparisons were done by imposing the differences between the mean of each group and the mean of combined groups. Also,

in Kruskal-Wallis test, the comparisons were done by imposing the differences between the rank of each group and the rank of combined groups. Yet such differences were also imposed in the power analysis of tests that are made up of a collection of paired comparisons between all possible pairs of groups such as T_1 and S_1 tests proposed by Babu *et al.* (1999). In other words, one single effect size index was usually used in power analysis to reflect the differences between treatment groups even in the tests that are based on a collection of paired comparisons between treatment groups. The latter type of tests should require more than one effect sizes to reflect the differences of the paired comparisons.

1.2 Pseudo-median procedure

The pseudo-median procedure is an alternative method for treatment groups comparison (Steland *et al.*, 2011), developed to deal with the problem of violation of the assumptions. It is a modification of the one-sample nonparametric Wilcoxon procedure in a two groups setting and extended to more than two groups. This procedure is made up of multiple pairwise comparisons between the control group and each of the other groups. The pseudo-median parameter is adopted as a location parameter to compare the treatment groups in the pseudo-median procedure. The pseudo-median parameter is the median of the distribution of the averages $(X_1 + X_2)/2$ where X_1 and X_2 are independently and identically distributed. This parameter is estimated by Hodges and Lehmann estimator.

1.3 Rationale of the study

The pseudo-median estimator has some characteristics which are better than the common robust estimators such as trimmed mean, *M*-estimator and median. The pseudo-median does not need to discard the data during the computation. However, the trimmed mean and *M*-estimator involve discarding some of the data that leads to loss of information. Moreover, the proportion of discards increases according to the number of extreme values. Consequently, more information will be lost when more data are discarded. At the same time, these extreme values represent part of the population with particular characteristics. Thus, the conclusion from previous procedures does not involve the analysis of this part of the population.

One other advantage of the pseudo-median is it exhibits good performance with the bootstrap method (Ahad *et al.*, 2011). Yet, the performance of the bootstrap on the sampling distribution of the sample median is very poor (Brown *et al.*, 2001).

Steland *et al.* (2011) proposed the pseudo-median procedure for comparing two groups and extended it to more than two groups. They estimated the power only at one point using two effect size values to reflect the differences between three groups, and concluded that the test is very reliable. This technique of obtaining power is deficient. This is because the performance of the method through only one point could not be determined.

Even though the pseudo-median has good characteristics, the power analysis results provided by Steland *et al*. (2011) was inconclusive. Estimating power at one point does not give correct and complete perception of the power performance. Moreover, the

manner of the pseudo-median procedure for comparing treatment groups depends on a group of multiple comparisons between the control and each of the treatment groups. This type of comparisons involves many different cases which should be considered when performing the power analysis.

Furthermore, developed tests which replaced the usual mean with other robust estimators did not yield high power under all conditions of heterogeneity and non-normality (Babu & Padmanabhan, 2002; Babu *et al.*, 1999; Keselman *et al.*, 2004b; Othman *et al.*, 2004a). Therefore, when using an improper method for power analysis, high and poor power situations could not be distinguish. Determining the poor power situations is important to treat performance problems in a test. The estimator used in a test could be a robust estimator but the methodology or the manner of the test for comparing groups leads to poor power.

1.4 Objective of the study

This research aims to develop a new power analysis technique for the pseudo-median procedure in order to measure the performance of the procedure in terms of Type I error and power when the assumptions of normality and homogeneity are violated.

The sub-objectives are as follows:

1. To establish Type I error of the pseudo-median procedure under various conditions.
2. To find suitable effect size estimator for the pseudo-median procedure.

3. To establish a new power analysis of the pseudo-median procedure and estimate a complete power curve under various conditions.
4. To compare the performances of the pseudo-median procedure against the classical parametric test (ANOVA *F*-test) and nonparametric test (Kruskal-Wallis test).

1.5 Significance of the study

This study contributes to an alternative power analysis of treatment group comparisons tests. The manner for comparing treatment groups is different from test to test which implies using different power analyses. The power of tests which depends on a collection of pairwise comparisons between the treatment groups could not be estimated at only a few points and with a single effect size. This study gives a new power analysis technique suitable for the tests which depend on multiple comparisons between the treatment groups.

1.6 Organization of the thesis

This thesis has five main chapters. Chapter 2 gives the gradual development in the area of comparing group of treatments and explains the main concepts related to the study such as Type I error, power, effect size, measures of robustness, pseudo-median parameter and its estimator and bootstrap. Chapter 3 describes the pseudo-median procedure, estimation of effect size and the technique of power analysis, study conditions for the simulations and the algorithms to calculate and evaluate performance

measures in the form of Type I error and power. The algorithms for computing the Type I error and power for the pseudo-median using bootstrap and the competing tests are also discussed in this chapter. Chapter 4 presents and illustrates the behavior of the pseudo-median procedure in terms of Type I error and power. Performances of the pseudo-median with the classical tests are also compared. Chapter 5 provides conclusions and suggestions for further studies.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

The comparisons of measurable quantities of characteristics of two or more groups are usually done in many scientific studies. The most used summary for measurable quantities is the mean. Comparing groups using means is the most common technique in education and psychology (Wilcox, 1995, p. 51). The analysis of variance is most widely used to compare means, specifically means of three or more groups. This classical test requires several assumptions to produce accurate results. The normality and homogeneity of data are the usual assumptions required for parametric tests. However, failures in the assumptions lead to distortion of Type I error and substantial reduction in the power of the test. Moreover, the assumptions are rarely met in real data. This chapter discusses the problem of violation of assumptions and reviews some of the solutions from previous studies. Also, it provides the definitions of terminologies related to this study.

2.2 Development of comparison of treatment groups

Numerous research have shown how the violations of the assumptions distort the Type I error and the power of a test. For example, Wilcox and Keselman (2003a) showed that the sampling distribution of a parametric test statistic departed from the true distribution when the observations were sampled from skewed distributions. This departure in the sampling distribution of the test statistic produced inaccurate Type I

error and confidence interval of the hypothesized parameter. The problem of mis-control over the probability of Type I error could be reduced when the sample size increases. Nevertheless, it persists even when the sample size is as large as n =300 with the presence of outliers (Wilcox & Keselman, 2003a). According to Wilcox (1994), the conventional F test failed to control the Type I error when the distributions were heterogeneous and/or non-normal. At $\alpha = 0.05$, the Type I error was greater than 0.09 with unequal variances and it exceeded 0.3 when the distributions were not normal as well.

Even when the data distributions are symmetric but not normal or slightly departing from normality, the power to detect the differences between group means was substantially reduced (Wilcox & Keselman, 2003a). For instance, in the two samples case, at $\alpha = 0.05$ and the variances were equal, the power of the Student's *t*-test was observed to be 0.28 falling from 0.975 with small departure from normality (Wilcox, 1995; Wilcox & Keselman, 2003a). This is even more when there is a large departure. The presence of outliers and heavy-tailed distributions result in the inflation of group sample variances which leads to lower power (Wilcox, 1995; Wilcox & Keselman, 2003a). For example, the power of the classical F test was reduced from 0.94 to 0.502 with symmetric heavy-tailed distributions and further reduced to 0.216 with skewed heavy-tailed distributions (Wilcox, 1994).

With inaccurate Type I error and poor power due to violation of normality and homogeneity assumptions, the decision of a test of hypothesis using parametric methods will be misleading. Hence, there is a danger in using parametric tests in the real world. This is because data in the real world is not normal and not homogeneous. According to

Reed (1998, p. 651) "*Nearly all real data are discrete in nature, so theory suggests that cannot be normal*". In addition, the variable reaction time, which is widely used in psychology and related fields is usually skewed (Miller, 1988). Micceri (1989) conducted surveys of 440 large samples to determine the properties of distributions that commonly occur in the real world. The study included a wide variety of measurements used in psychology and education (e.g., psychometric measures, ability and aptitude measures). None of these data had normal distributions and few of them were approximately normal in shape. Furthermore, most of the data classified were skewed, extremely skewed, heavy-tailed and multimodal.

With regards to heterogeneity, this phenomenon in real data is not something strange. Given the nature of research, as well as the populations from which samples were drawn heterogeneity is common. O'Brien (1992, p. 819) noted that when comparing patients who have a certain disease with non-infected patients, there were variability in the laboratory measures in both groups. Irregular behavior resulting from the impact of certain treatments can also cause more variability (Steel & Torrie, 1981 pp. 169-170). In addition, in studies of psychology there are variables that naturally showed heterogeneity within groups that share a common trait. One such variable is reaction time, e.g., heterogeneity of reaction times among age groups (Hultsch *et al.*, 2002), gender and education level.

Both assumptions (normality and heterogeneity) can be tested statistically. However, the methods used to detect them also require assumptions. The methods for detecting the equality of variances require normality and normality tests require homogeneity (Erceg-Hurn & Mirosevich, 2008; Montgomery, 2001). Therefore, neglecting the normality

condition in a test of equality of variances or absence of homogeneity in a test of normality gives incorrect decision. Furthermore, most methods which test homogeneity cannot control Type I error and have low power (Wilcox, 1995). Moreover, the common normality tests, such as the chi-squared test in goodness-of-fit setting and the Kolmogorov-Smirnov test give poor power and should not be used for testing normality (D' Agostino *et al.*, 1990).

Data transformation is one of the solutions to obtain normal and/or homogeneous data. Yet, the researchers find difficulty in interpreting the results because the data unit after transformation is different from the original data. Also, determining the appropriate transformation to deal with both heterogeneity and non-normality is not easy. Furthermore, the outliers are not necessarily treated or removed using the transformations (Erceg-Hurn & Mirosevich, 2008, p. 594; Keselman *et al.*, 2007, p. 269; Wilcox, 1995, pp. 69-70).

Nonparametric methods are also used when there are violations in the assumptions. Usually, nonparametric methods can detect the difference between group treatments under non-normal symmetric distributions. However, these methods are still affected by the heterogeneity condition, e.g. the Kruskal-Wallis procedure is affected by heterogeneity whether the design is balanced or unbalanced. Furthermore, nonparametric methods usually have less power than parametric methods and need larger sample sizes to reject false hypotheses (Keselman *et al.*, 2007, pp. 268-269; Syed Yahaya *et al.*, 2006, p. 50).

Certain research turned to replacing the usual least squares estimators (the usual mean and variance) with other estimators which are less sensitive to non-normality and heterogeneity. The frequent location estimators that were adopted and examined instead of the usual mean are the sample median, trimmed mean and the *M*-estimator. These estimators achieved good Type I error in some studies. Yet, they have weaknesses especially in extremely skewed distributions (Babu *et al.*, 1999; Lix & Keselman, 1998; Wilcox & Keselman, 2003b; Wilcox *et al.*, 1998).

Lix and Kesleman (1998) examined the ANOVA *F*-test and other alternative procedures such as Welch (1951), Alexander and Govern (1994) and Box (1954) were compared when the underlying distributions were non-normal and also the group variances and sample sizes were jointly unequal. They employed the trimmed means and Winsorized variances instead of least square estimators (the usual means and variances) in all test statistics that were adopted. They recommended that using trimmed means and Winsorized variances achieve good control of the Type I error and high rate of power in some of the alternatives. At the same time, Wilcox *et al.* (1998) examined the methods due to Welch (1951), Alexander and Govern (1994) and Box (1954). They also used the trimmed means and Winsorized variances as well as applied bootstrap under the same assumptions in Lix and Keselman (1998). They showed that better control of Type I error can be obtained if the bootstrap method is used in conjunction with test statistics based on trimmed mean. Keselman *et al.* (2000) examined two procedures for equality of means proposed by Weerahandi (1995) and Chen and Chen (1998). They also compared these two procedures with robust Welch test with 20% trimmed means and Winsorized variances examined by Lix and Keselman (1998). They showed that under

normality, these procedures were robust when the group variances were heteroscedastic and the sample sizes were unequal. However, they provided large Type I error when the data were not normal. In contrast, the robust Welch test provided better control of Type I error in similar situations.

Subsequently, Wilcox and Keselman (2003b) illustrated the concerns of the trimmed mean. These concerns are: 1) the amount of trimming has to be fixed before analysis of data and 2) the nature of trimming, whether symmetric or asymmetric. Usually trimming is carried out symmetrically regardless of whether the distribution is symmetric or skewed. Therefore, they modified the *M*-estimator into a one-step *M*-estimator (*MOM*). *MOM* is also a robust estimator which simultaneously controls the trend and the magnitude of the necessary trimming. They demonstrated that the *MOM* estimator was able to control the Type I error better than the trimmed mean.

A further evolution to address the concerns of the amount of trimming, Tukey and McLaughlin (1963), Jaeckel (1972) and Hogg *et al.* (1975) proposed a method to determine the magnitude of trimming. They suggested choosing the strategy which results in the smallest standard deviation of the sample trimmed mean. In other words, they computed many different trimmed means and then adopted the one which has the smallest standard deviation. Subsequently, Reed and Stark (1996) developed adaptive location estimators based on measure of tail length and measure of skewness for a group of *n* observations. For this adaptive estimator, the amount and the trend of trimming either symmetrically or asymmetrically is determined by the characteristics of the sample data such as the tail length and the degree of skewness. Following that, Keselman *et al.* (2007) applied the adaptive trimmed means with the Welch (1951)

statistic using Tukey-McLaughlin-Jaeckel-Hogg methods and Reed and Stark estimators. They found that a number of Welch tests based on Reed and Stark estimators provided good values of Type I error under less extreme cases of non-normality and variance heterogeneity.

As stated earlier about the adaptation, Babu *et al.* (1999) proposed adaptive method for treatment groups comparison using a different manner of adaptation. They introduced two tests statistics, T_1 and S_1. The T_1 is based on 15%-trimmed means while S_1 is based on sample medians. The strategy of this adaptive method is to start first with checking the data by using preliminary test for symmetry in each simulation. If the data is symmetric, the T_1 statistics is used; otherwise, the S_1 statistics is used.

Consequently, Keselman *et al.* (2002) and Othman *et al.* (2004b) employed another adaptive method. They used Babu *et al*. (1999) test for symmetry to trim symmetrically or asymmetrically only on one side. Once the data have been symmetrically or asymmetrically trimmed, a number of Welch-James heteroscedastic statistics were calculated. The Welch-James heteroscedastic statistics are the Welch (1951) test after replacing the usual mean and variance by trimmed means and Winsorized variances with different α% trimming. The Welch-James heteroscedastic statistics are transformed using both Johnson's (1978) or Hall's (1992) transformation with or without employing bootstrap to calculate the empirical values of the Type I error. Their results showed good control of the Type I error when the Welch-James heteroscedastic statistic is preceded by the Babu et al. (1999) test for symmetry. Then, followed by 10% symmetrically trimmed or 20% asymmetrically trimmed means with either Johnson's (1978) or Hall's (1992) transformation in conjunction with the bootstrap method.

Concerning other developments in treatment groups comparison, Md. Yusof *et al.* (2008) modified the T_1 statistic proposed by Babu *et al.* (1999) by using variable trimmed mean and Winsorized variances based upon several robust scale estimators in the trimming criterion. They showed, in general, the original T_1 procedure from Babu *et al.* (1999) is still the best. Nevertheless, the methods using the scale estimators improved the Type I error rate when the sample size was large.

Regarding comparison of group medians, the S_1 statistic which was proposed by Babu *et al.* (1999) for comparing group medians was modified by replacing the default standard error of the sample median, $\hat{\omega}$, in the S_1 statistics with alternative robust scale estimators proposed by Rousseeuw and Croux (1993) (Othman *et al.*, 2006; Syed Yahaya *et al.*, 2004a, 2004b; Yaacob *et al.*, 2006). Some of these alternative robust scale estimators when combined with S_1 statistic achieved good control of the Type I error and high power. The two robust scale estimators, MAD_n and T_n, achieved the best control of the Type I error with S_1 statistic compared to the other robust scale estimators (Othman *et al.*, 2006; Syed Yahaya *et al.*, 2004a, 2004b)

A further development in comparing location parameters, Keselman *et al.* (2002) created a new procedure by applying the *MOM* estimator on *H* statistic which was due to Schrader and Hettmansperger (1980). They called this the *MOM-H*. The bootstrap procedure was used to determine the critical value of *MOM-H*. Subsequently, frequent investigations were done on *MOM-H* (Othman *et al.*, 2006; Syed Yahaya *et al.*, 2006; Yaacob *et al.*, 2006). All investigations involved modifying the trimming criterion by replacing the default scale estimator (MAD_n) with other robust scale estimators suggested by Rousseeuw and Croux (1993). They showed that the T_n and S_n robust scale

estimators achieved the best performance with *MOM-H* when the data was non-normal and heteroscedastic (Othman *et al.*, 2004a; Othman *et al.*, 2006; Syed Yahaya *et al.*, 2006). More scale estimators, specifically E_1 and E_2 were used by Yaacob *et al*. (2006). They did not show better control of Type I error than the default estimator MAD_n.

In addition to the robust procedures mentioned earlier, there were developments in using the Mann-Whitney statistic to compare more than two groups. Prior to application of more than two groups, this statistic has to be fixed to become applicable to non-symmetric distributions. Babu and Padmanabhan (2002) tried to improve the Mann-Whitney procedure to make it applicable for skewed distributions. The Mann-Whitney criterion, $P(X \leq Y) = 0.5$, cannot be used when the distributions are asymmetric and the variances are unequal. Therefore, they modified the Mann-Whitney procedure by estimating the probability $P(X \leq Y)$. This probability was estimated by employing the bootstrap method. Their procedure resulted in poor performance of the Type I error and power. At the same time, Othman *et al*. (2003) extended the Mann-Whitney to *J*-samples, where $J > 2$ using the same procedure. They obtained liberal rates of the Type I error similar to the results in Babu and Padmanabhan (2002) especially when the variances of the groups were extremely different.

2.3 Type I error

The Type I error is one of the fundamental concepts of tests of hypothesis. The statisticians defined it as the probability of rejecting the null hypothesis when it is true. The significance level and the error of the first kind are various names for the Type I

error. The size of Type I error is denoted by α (Cohen, 1988). Usually the value of α is specified in the test of hypothesis. Practitioners commonly choose α to be 0.01, 0.05 and 0.1, and 0.05 is the most frequently used (Cohen, 1994; Cowles & Davis, 1982).

When the assumptions of normality and heterogeneity are verified, the probability of the Type I error for parametric tests are usually close to the set level α. However, the probability departs from the nominal significance level α when the assumptions are violated. A robust statistic is a procedure which is able to maintain the Type I error close to the nominal level and maintain the power when the assumptions are violated (Stevens, 2007).

Bradley (1978) considered that a test to be robust if the departure of the probability of the Type I error, *p*, from the nominal level α was within the interval $0.5\alpha \leq p \leq 1.5\alpha$. This criterion of robustness is called the liberal criterion and is widely used in numerous researches (Keselman *et al.*, 2007; Keselman *et al.*, 2000; Othman *et al.*, 2004a; Wilcox, 1994; Wilcox & Keselman, 2003b; Wilcox *et al.*, 1998). If the nominal level is set at $\alpha = 0.05$, the liberal criterion will be [0.025, 0.075]. Type I error above 0.075 is considered liberal and if below 0.025, it is considered conservative.

Some researchers used different criterion of robustness. They used the confidence interval of the proportion, *p*, $(\hat{p} \pm z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n})$ (Babu *et al.*, 1999; Syed Yahaya *et al.*, 2004b). The bounds of the interval were computed by setting $\hat{p}$ equal to α and *n* is the number of simulations while $z_{\alpha/2}$ is the critical value from the standard normal table. The dependence of this criterion on the number of simulations made the criterion more accurate since the larger the *n* is, the smaller the interval becomes. When $\alpha = 0.05$

and $n = 5000$, the accurate criterion will be [0.044, 0.056] (Babu *et al.*, 1999; Syed Yahaya *et al.*, 2004b)

2.4 Power

Besides Type I error, the power of a statistical test is one of the performance metrics that distinguishes a test from another. When there exists more than one statistical test for testing a specific problem, the power of these procedures leads to the determination of which test is best to use. Practitioners prefer the statistical test which has high power (Mahoney & Magel, 1996).

Power is defined as the probability of rejecting the null hypothesis given that the alternative hypothesis is true, or in other words, the probability that leads to significant results. Computing this probability needs complete knowledge of the population distribution (Mahoney & Magel, 1996). This requirement makes the power computation process difficult especially for nonparametric methods that are applied when the population distributions are unknown. Therefore, researchers sometimes pretend that the power cannot be assessed without empirical data. On the other hand, Cohen (1988) illustrated that the power analysis depends upon three components: the size of the test or "significance level", the sample size and the effect size. Power is obtained at a specific value of the effect size, when the sample size and the significance level are fixed.

2.5 Effect size

The effect size is a measure to quantify the degree of departure of the decision from the null hypothesis. For example, in a study to determine whether the average scores of students in class A, μ_A, is different from the average scores in class B, μ_B, the null hypothesis is $\mu_A - \mu_B = 0$. This implies no difference between the two means. If the difference is a specific nonzero value, the effect size is this specific value.

The procedure to obtain the effect size differs from one statistical test to another. Each statistical test has its own effect size index (Cohen, 1988). Hence, parametric statistical tests have parametric effect sizes. One such effect size is the *d* index, an effect size for the difference between two population means in independent samples *t* test (Cohen, 1992). Another effect size is the f index, the effect size for equality of a set of *k* population means in the analysis of variance procedure (Cohen, 1988). Both of these indices are constructed from specific values of the alternative hypothesis, sample size and the size of Type I error. However, these are effect sizes from parametric statistics. They are affected by departures from normality and homogeneity (Algina *et al.*, 2005; Erceg-Hurn & Mirosevich, 2008; Hogarty & Kromrey, 2001; Onwuegbuzie & Daniel, 2002). Therefore, researchers need to find nonparametric effect sizes that are applicable for nonparametric statistics.

A number of researches have been done to develop nonparametric effect size. Cliff (1993) proposed the delta statistic denoted by δ as a non-parametric index to quantify the differences between two groups on ordinal level measurements. The delta statistics is the difference between the probability of a score from the first group being larger than

the second group minus the probability of the second group being larger than the first group.

McGraw and Wong (1992) developed a Common Language (*CL*) effect size index for continuous distributions. The *CL* presents the effect size in probability which is "*the probability that a score sampled at random from one distribution will be greater than a score sampled at random from some other distribution*" (McGraw & Wong, 1992, p. 361). Also, the *CL* was generalized to independent and correlated *n*-groups.

Vargha and Delaney (2000) modified the *CL* index to be applicable for any discrete or continuous variable. It is called the *measure of stochastic superiority* and is denoted by *A*. This effect size does not require conditions about the type of distribution. It only requires that the distribution to be at least ordinally scaled. The *A* index converts the effect size into a probability. Vargha and Delaney gave guidelines for interpreting the value of *A*. The value $A = 0.5$ indicates equality of two populations while $A > 0.5$ means that the first population is superior to the second population. Vargha and Delaney gave three levels of the effect size *A*. The value $A = 0.56$ is considered as a small effect size, $A = 0.64$ as a medium while $A = 0.71$ as a large effect size.

Vargha and Delaney found a relationship between *A* and δ which is $A = (\delta + 1)/2$. Both effect sizes were demonstrated among other effect sizes to be robust in violation of the assumptions of normality and homogeneity (Hogarty & Kromrey, 2001; Leech & Onwuegbuzie, 2002).

The effect size plays an important role in the calculation of power. The *f* effect size index is usually employed in power analysis of treatment group comparisons to reflect the differences between groups (Keselman *et al.*, 2004b; Othman *et al.*, 2004a).

Cohen (1988) gave three degrees for the *f* effect size index. The value $f = 0.10$ is considered as a small effect size, $f = 0.25$ as a medium while $f = 0.40$ as a large effect size. Three patterns of variability, minimum, intermediate and maximum variability were also defined to reflect how means of groups deviate from each other. The patterns are functions in *f*, and each pattern indicates one degree of *f* effect size.

Keselman *et al.* (2004b) and Othman *et al.* (2004a) used these patterns in the power analysis for *MOM-H* and *MOM-T* procedures by Keselman *et al.* (2002), and the adaptive procedure by Babu *et al.* (1999). The *MOM-H* procedure is made up of comparisons between each treatment group and the combined groups while the other two procedures are made up of multiple comparisons between all possible pairs of groups. The formula of the *f* effect size is based on differences between the mean of each treatment group and the mean of the combined groups. Also, the *f* effect size or the patterns do not express how many paired comparisons are different or what the degree of effect size is in each paired comparison. These matters need more than one effect size to obtain a complete picture for power performance of the procedure.

Steland *et al.* (2011) and Babu *et al.* (1999) proposed two procedures made up of multiple comparisons between groups. Three groups were considered for power analysis for these two procedures. Two non-zero effect size values were used to present the location shift in the second and third groups, respectively. This implies that all groups

are different. However, this is not the only case of differences between the groups which could occur.

2.6 Measures of robustness

Researchers have been trying to find robust estimators which are less sensitive to small deviations from the usual assumptions. The influence function and the breakdown points are tools to describe and measure the stability or robustness of the statistics. The influence function describes the limiting effect of an additional observation, *x*, to a very large sample on a statistic *T* (Hampel *et al.*, 1986; Wilcox, 2005). In other words, the influence function reflects the approximate rate of change of the estimate when the outlier occurs (Hettmansperger & McKean, 1998). Limited change on a statistic by an additional value leads to a resistant or stable estimator. Therefore, a bounded influence function leads to a robust estimator.

The other measure of robustness is the breakdown point which reflects the amount of contaminated data that an estimator can cope with (Hettmansperger, 1984; Huber, 1981). The estimator with high breakdown point is considered resistant and robust. A high breakdown point is one of the characteristics of a robust estimator. The sample mean has 0 breakdown point while the α-trimmed mean has α breakdown point. The sample median has a high breakdown point equal to 0.5.

2.7 Pseudo-median parameter and its estimator

The pseudo-median is a measure of location which is used in the pseudo-median statistical procedure to compare a group of treatments. Høyland (1965, p. 178) defined the pseudo-median *"of a distribution F as the median of the distribution of* $(X_1 + X_2)/2$ *where* X_1 *and* X_2 *are independently and identically distributed according to F"*. The median and the pseudo-median are identical when *F* is symmetric (Høyland, 1965).

The consistent estimator for the pseudo-median parameter is the Hodges–Lehmann estimator denoted by *HL*. There are different types of Hodges–Lehmann estimators for one and two samples problem (Hodges & Lehmann, 1963). These estimators measure the location difference of two samples (Everitt, 2006). For one sample, *HL* estimator is considered a corresponding estimator of the pseudo-median parameter, θ, and it is given by

$$HL = \text{median}\left\{\frac{x_i + x_j}{2}, i \leq j = 1,2, \dots, n\right\} \tag{2.1}$$

where *n* is the sample size and x_i and x_j refer to the observations. For two samples, x_i and x_j are replaced by d_i and d_j where d_i and d_j are the differences between the observations of the two samples.

HL statistics has a number of advantages. One of the advantages is this statistic can be used in regression and generalized to multivariate statistics (Hettmansperger & McKean, 1998; Oja, 2010) and other areas of statistics depending on the rank or sign rank, such as Wilcoxon sign rank test and Wilcoxon rank sum test (Hollander & Wolfe, 1999, p. 54 and p. 126). In addition, the *HL* estimator has some properties of a robust

estimator. It is insensitive to outliers (Hollander & Wolfe, 1999; Lehmann, 2006). Furthermore, it has bounded influence function. Also, its breakdown point is 0.29. However, the breakdown point of the trimmed mean is α which is the percentage of trimming. Usually this percentage does not exceed 20%; otherwise, more information will be lost.

2.8 The Bootstrap

Efron (1979) introduced the bootstrap for estimating the standard error of an estimator. The bootstrap is a resampling technique from the original data set used to obtain a pseudo sampling distribution of a statistic. It replaces the theoretical distribution of a statistic by an empirical one when the theoretical distribution of a statistic is complicated or unknown. This technique is a practical and simple way to estimate the properties of an estimator and in constructing a test involving the same estimator. The bootstrap technique is also used to provide an approximate sampling distribution when the usual assumptions are not satisfied or when the standard error of a statistics has a complex formula. In robust statistics, many studies demonstrated that good results of the Type I error were obtained when combining the bootstrap method with statistical procedures based on robust estimator (Keselman *et al.*, 2002; Othman *et al.*, 2004b; Wilcox, 1995; Wilcox *et al.*, 1998).