ROBUST PERCENTILE BOOTSTRAP TEST WITH MODIFIED ONE-STEP *M*-ESTIMATOR (*MOM*): AN ALTERNATIVE MODERN STATISTICAL ANALYSIS

NURUL HANIS BINTI HARUN

MASTER OF SCIENCE (STATISTICS)
UNIVERSITI UTARA MALAYSIA
2015

Awang Had Salleh
Graduate School
of Arts And Sciences

Universiti Utara Malaysia

PERAKUAN KERJA TESIS / DISERTASI
(Certification of thesis / dissertation)

Kami, yang bertandatangan, memperakukan bahawa
(We, the undersigned, certify that)

NURUL HANIS HARUN

calon untuk Ijazah
(candidate for the degree of) MASTER

telah mengemukakan tesis / disertasi yang bertajuk:
(has presented his/her thesis / dissertation of the following title):

"ROBUST PERCENTILE BOOTSTRAP TEST WITH MODIFIED ONE-STEP M-ESTIMATOR (MOM): AN ALTERNATIVE MODERN STATISTICAL ANALYSIS"

seperti yang tercatat di muka surat tajuk dan kulit tesis / disertasi.
(as it appears on the title page and front cover of the thesis / dissertation).

Bahawa tesis/disertasi tersebut boleh diterima dari segi bentuk serta kandungan dan meliputi bidang ilmu dengan memuaskan, sebagaimana yang ditunjukkan oleh calon dalam ujian lisan yang diadakan pada : ***29 September 2014.***
That the said thesis/dissertation is acceptable in form and content and displays a satisfactory knowledge of the field of study as demonstrated by the candidate through an oral examination held on: ***September 29, 2014.***

Pengerusi Viva: *(Chairman for VIVA)*	**Assoc. Prof. Dr. Haslinda Ibrahim**	Tandatangan *(Signature)*
Pemeriksa Luar: *(External Examiner)*	**Dr. Norhashidah Awang**	Tandatangan *(Signature)*
Pemeriksa Dalam: *(Internal Examiner)*	**Assoc. Prof. Dr. Sharipah Soaad Syed Yahaya**	Tandatangan *(Signature)*
Nama Penyelia/Penyelia-penyelia: *(Name of Supervisor/Supervisors)*	**Dr. Zahayu Md Yusof**	Tandatangan *(Signature)*

Tarikh:
(Date) ***September 29, 2014***

Permission to Use

In presenting this thesis in fulfillment 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 copying of this thesis in any manner, in whole or in part, for scholarly purpose may be granted by my supervisor or, in her 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

Kenormalan dan homoskedastisiti merupakan dua andaian utama yang perlu dipenuhi apabila berurusan dengan ujian-ujian parameter klasik untuk perbandingan kumpulan. Pelanggaran mana-mana andaian tersebut akan menyebabkan keputusan ujian menjadi tidak sah. Walau bagaimanapun, pada realitinya, kedua-dua andaian tersebut sukar dicapai. Untuk mengatasi masalah tersebut, kajian ini mencadangkan pengubahsuaian satu kaedah yang dikenali sebagai ujian Bootstrap Berparameter dengan menggantikan min sebenar, $\bar{X}$ dengan ukuran lokasi yang sangat teguh iaitu penganggar-M satu-langkah terubahsuai (*MOM*). (*MOM*) merupakan min terpangkas tidak simetri. Penggantian ini akan menjadikan ujian Bootstrap Berparameter lebih teguh untuk perbandingan kumpulan. Dalam kajian ini, kriteria pemangkasan untuk *MOM* menggunakan dua pengganggar skala yang amat teguh iaitu MAD_n dan T_n. Satu kajian simulasi telah dijalankan untuk mengkaji prestasi kaedah yang dicadangkan berdasarkan kadar Ralat Jenis I. Untuk mengenal pasti kekuatan dan kelemahan kaedah, lima pembolehubah iaitu: bilangan kumpulan, saiz sampel seimbang dan tak seimbang, jenis taburan, keheterogenan varians, dan sifat pasangan bagi saiz sampel dan varians kumpulan dimanipulasi untuk menghasilkan pelbagai keadaan yang biasanya wujud dalam kehidupan sebenar. Prestasi kaedah yang dicadangkan kemudiannya dibandingkan dengan ujian parameter klasik dan ujian tidak berparameter yang paling kerap digunakan untuk dua (ujian-*t* tidak bersandar dan ujian *Mann Whitney* masing-masing) dan lebih daripada dua kumpulan tidak bersandar (ANOVA dan ujian *Kruskal Wallis* masing-masing). Dapatan kajian menunjukkan bahawa, untuk dua kumpulan, ujian Bootstrap Berparameter yang teguh menunjukkan prestasi yang baik di bawah keadaan varians heterogen dengan taburan normal atau taburan terpencong. Manakala untuk lebih daripada dua kumpulan, ujian tersebut menjana pengawalan Ralat Jenis I yang baik di bawah varians heterogen dan taburan terpencong. Dalam perbandingan dengan kaedah parameter klasik dan keadah tidak berparameter, ujian yang dicadangkan menunjukkan prestasi yang lebih baik di bawah taburan terpencong dan varians heterogen. Prestasi setiap prosedur juga ditunjukkan dengan menggunakan data sebenar. Secara umumnya, prestasi Ralat Jenis I bagi ujian yang dicadangkan adalah sangat menyakinkan walaupun andaian kenormalan dan homoskedastisiti dilanggar.

Kata kunci: Titik kegagalan, Heterogen, Taburan terpencong, Ralat Jenis I.

Abstract

Normality and homoscedasticity are two main assumptions that must be fulfilled when dealing with classical parametric tests for comparing groups. Any violation of the assumptions will cause the results to be invalid. However, in reality, these assumptions are hardly achieved. To overcome such problem, this study proposed to modify a method known as Parametric Bootstrap test by substituting the usual mean, $\bar{X}$ with a highly robust location measure, modified one step M-estimator (*MOM*). *MOM* is an asymmetric trimmed mean. The substitution will make the Parametric Bootstrap test more robust for comparing groups. For this study, the trimming criteria for *MOM* employed two highly robust scale estimators namely MAD_n and T_n. A simulation study was conducted to investigate on the performance of the proposed method based on Type I error rates. To highlight the strength and weakness of the method, five variables: number of groups, balanced and unbalanced sample sizes, types of distributions, variances heterogeneity and nature of pairings of sample sizes and group variances were manipulated to create various conditions which are common to real life situations.The performance of the proposed method was then compared with the most frequently used parametric and non parametric tests for two (independent sample *t*-test and Mann Whitney respectively) and more than two independent groups (ANOVA and Kruskal Wallis respectively). The finding of this study indicated that, for two groups, the robust Parametric Bootstrap test performed reasonably well under the conditions of heterogeneous variances with normal or skewed distributions. While for more than two groups, the test generate good Type I error control under heterogeneous variances and skewed distributions. In comparison with the parametric and non parametric methods, the proposed test outperforms its counterparts under non-normal distribution and heterogeneous variances. The performance of each procedure was also demonstrated using real data. In general, the performance of Type I error for the proposed test is very convincing even when the assumptions of normality and homoscedasticity are violated.

Keywords: Breakdown point, Heterogeneity, Skewed distributions, Type I error.

Acknowledgement

This thesis would not been possible without guidance and help of several individuals who contributed their assistance in the preparation and completion of this study. It gives me great pleasure to acknowledge their support.

First and foremost, I would like to express the deepest appreciation and gratitude to my supervisor, Dr. Zahayu Md Yusof for her valuable support and guidance throughout this study. I could not have imagined having a better advisor and supporter for my master study in University Utara Malaysia.

I am deeply grateful to my parents, Harun Hamid and Sharifah Abu Bakar for their love, inspiration, patience and support. I dedicated this work to my parents. I would also like to thank to my aunt and my brother for their support.

Lastly, I would like to thank everybody who had directly or indirectly helped me during this research.

Table of Contents

List of Tables

List of Figures

List of Appendices

List of Abbreviations

ANOVA	Analysis of variance
Parametric Bootstrap	A statistical method for testing the equality of central tendency
$\boldsymbol{MAD_n}$	Median absolute deviation about median
$\boldsymbol{T_n}$	A scale estimator

List of Publications

Md Yusof, Z., Harun, N. H., Syed Yahaya, S. S. & Abdullah, S. (2013). A modified parametric bootstrap: an alternative to classical parametric test. *In proceeding of the World Conference on Integration of Knowledge 2013*, 25 – 26 November, Langkawi, Malaysia.

Harun, N. H, Md Yusof, Z. (2013). Testing the Equality of Central Tendency using Robust Parametric Bootstrap Test with *MOM* Estimator for Two Groups Case. *In proceeding of the 1st Innovation and Analytics Conference and Exhibition 2013*, 29 December, Universiti Utara Malaysia, Malaysia.

Harun, N. H, Md Yusof, Z. (2014). Robust Parametric Bootstrap Test with *MOM* Estimator: An Alternative to Independent Sample *t*-Test. In proceeding of the *3rd International Conference on Quantitative Sciences and Its Applications 2014*, 12 – 14 August, Langkawi, Malaysia.

CHAPTER ONE

BACKGROUND

1.1 Introduction

Statistics encompasses a wide variety of activities, ideas and results that can handle the situations involving uncertainties. Statistics consists of two basic statistical analysis namely descriptive statistics and inferential statistics. Recording and summarizing a data set is the main purpose of descriptive statistics whereas inferential statistics involves drawing conclusion and making decisions. There are extensive studies in testing equality of central tendency measures in inferential statistics using statistical method in order to make inferences based on obtained results. Basically, classical parametric tests such as analysis of variance (ANOVA) and independent sample *t*-test are often used in testing the central tendency measure by researchers rather than other methods since the aforementioned methods provide a good control of Type I error and generally more powerful than other methods when all the assumptions are fulfilled (Wilcox & Keselman, 2010).

ANOVA is used to determine the mean equality for more than two groups while independent samples *t*-test is used to determine the mean equality for two independent groups. However, a characteristic of these procedures is the fact that making inference depends on certain assumptions that need to be fulfilled. There are three main assumptions that need to be fulfilled before making inference on the classical parametric test such as: (a) collecting data from independents groups, (b) normally distributed data and (c) variances in the groups are equal (homoscedasticity). However, the specific interest of this study is to focus only for

the assumptions of normality and equality of variances in the groups since these assumptions are rarely met in real data.

Normal distribution can be defined as a symmetrical distribution that describes the expected probability distribution of many chance occurrences. It forms a bell shape curve. Zikmund, Babin, Carr & Griffin (2010) reported it as one of the most common probability distribution of statistics. However, normality assumption can easily be violated if the distribution is skewed and in the presence of outliers in single data set. Outliers can be defined as unusually large or small value in a data set. Wilcox (2002) stated in his study that the hypotheses testing method based on the equality of central tendency such as mean can have poor properties (e.g. reduce statistical power and reduce the ability in controlling the Type I error) if skewness or outliers or both tend to appear in data set.

Another problem with classical parametric procedures occurs when the groups have unequal variances (heteroscedasticity). According to Wilcox and Keselman (2010), unequal variances can cause classical parametric test results to be biased even though all groups have normal distribution. Moreover, Kohr & Games (1974) pointed out that unequal variances in the groups can affect the validity and reliability of the classical parametric test especially for unbalanced sample sizes.

According to Erceg-Hurn and Mirosevich (2008), violation in the assumptions of normality and homogeneity of variances can have drastic effect on the result of classical parametric test especially on the Type I error and the Type II error. Type I error occur when the null hypothesis is rejected even though it is true while Type II error occur when the false null hypothesis is failed to reject.

Failure to meet the assumptions of normality and equality of variances can distort the Type I error rates. For example, the probability of Type I error must be within the level of significant bound when the null hypothesis is assumed to be true. However, violation in any of these assumptions can lead to inflated the Type I error rates and consequently will make the Type I error contained outside the level of significant bound when the null hypothesis is assumed to be true (Wilcox & Keselman, 2010). Therefore, the results produced by the test that is used may become invalid.

As mentioned earlier, both of classical test (e.g. ANOVA and independent sample *t*-test) are based on certain assumptions such as normality distribution and equality of variances. In real situation, data that fulfilled both assumptions is hard to find. Thus, a common recommendation is to use non-parametric test or simple transformation.

A distribution-free procedure (non-parametric) frequently used as an alternative since they are valid under very general assumptions. Mann Whitney test and Kruskal Wallis test can be alternative procedure to ANOVA and independent sample *t*-test, respectively. Non-parametric test is known as a quick procedure, simple and can be calculated by hand for small sample sizes.

However, non-parametric test is not without disadvantages. Although non-parametric statistics is currently one of the most important branches of statistics but they are criticized because of some reasons. Daniel (1990) stated that the arithmetic in many instances is tedious and laborious when sample sizes are large and a computer is not handy even though non-parametric test have a reputation for requiring only simple calculations. Apart from that, non-parametric procedures are not as powerful as

classical parametric test and require larger sample sizes to reject a false hypothesis (Syed Yahaya, Othman & Keselman, 2006).

Another alternative when dealing with non-normal distribution and unequal variances is transforming the data into another scale in the same manner by using simple transformations. Simple transformations methods that often use are logarithms, square roots or inverse transformation. However, simple transformations also have made some issue. Transformations data are failed to deal effectively with outliers even though they can alter distribution to make the data more symmetrical (Wilcox & Keselman, 2003). Sometimes, when using simple transformation, outliers still remain and this condition can reduce the statistical power. This is because simple transformations do not eliminate the effect of outliers. Besides, the value produced by simple transformations had made some issues in the interpretation of the data since it involves placing the data on another scale (Lix, Keselman & Keselman, 1996).

Hence, developing a test statistics which is appropriate under non-normal distribution and unequal variances became goal for researchers. Thus, robust statistical procedures have been identified as alternative procedures which have good control of Type I error rates even under non-normal distribution and heterogeneous variances. Robust estimator is stable and insensitive to all of these violations, which it can deal effectively with outliers and skewed distribution. In other words, robust test can provide a good control on Type I error even if there are skewness or outliers in a data set (Wilcox & Keselman, 2010; Othman, Keselman, Wilcox, Fradette & Padmanabhan, 2002; Lix & Keselman, 1998).

1.2 Robust Statistics

The term “robustness” was first used by Box (1953). His study introduced the need for robust method and at the same time can be seen as a breakthrough in robust statistics field. Among the earlier procedures used by researchers are Welch test (1951), James test (1951) and Box (1953). Then, Huber (1964) and Hampel (1974) did an extensive research regarding robust statistics. Since then, a lot of finding showed that robust test can be advantages compared to classical parametric test and non-parametric test when the assumptions of classical parametric test are violated (Md Yusof, Abdullah & Syed Yahaya, 2012a; Md Yusof, Harun, Syed Yahaya & Abdullah, 2013).

Many researchers tried to define robust statistics properly. According to Huber (1981), robustness signifies insensitivity to small deviations from the assumptions. He also stated that a model is considered robust if it is reasonably efficient, small deviations from the model assumptions will not drastically impair the performance of the model and somewhat large deviations from the model will not invalidate the model. Apart from that, according to Scheffe (1959), a statistical method is considered robust if the inferences are not seriously invalidated by the violation of normality distribution and equality of variances.

Classical parametric test is a powerful test that can give an accurate result. However, with the advances and insights achieved by researchers nowadays showed that violation in the assumptions of classical parametric test can lead to bias and distort the results. Hence, robust test uses the advantages of classical parametric models but allows violation in the assumptions of classical parametric test to maintain a good control of Type I error rates and the statistical power.

Robust procedures involve replacing the original mean and variances with robust measures of location and scale. For example, some researchers proposed using trimmed mean and Winsorized variances when applying alternatives approaches such as James test and Welch test since these robust procedures can improve robustness (Lix & Keselman, 1998: Keselman, Wilcox, Othman & Fradette, 2002). The applying test intended to provide better Type I error control when computed with trimmed means and Winsorized variances (Lix & Keselman, 1998).

Among the latest procedures in robust statistics is Parametric Bootstrap test with trimmed mean and Winsorized variances proposed by Cribbie, Fiksenbaum, Keselman and Wilcox (2012). The findings from their study indicated that Parametric Bootstrap test with trimmed means and Winsorized variances produced Type I error rates close to the nominal value of $\alpha = 0.05$ under the conditions of non-normal distribution and unequal variances. Therefore, the Parametric Bootstrap test will be the main focus in this study in order to investigate the performances of the proposed procedure in testing the equality of central tendency.

1.3 Parametric Bootstrap Test

Parametric Bootstrap test was originally introduced by Krishnamoorthy, Lu, and Mathew (2007) as a new statistical test for comparing the equality of central tendency measures such as means of independent groups under the presences of variances heterogeneity. The objective of Krishnamoorthy *et al.* (2007) study was to compare the performances of proposed Parametric Bootstrap test with Welch Test, James test and the generalized F (FG) test under unequal variances in the groups. The result showed that Parametric Bootstrap test has a good control of Type I error even for small sample sizes and the number of groups was large.

However, Krishnamoorthy *et al.* (2007) did not explore the performance of Parametric Bootstrap under the condition of non-normal distribution. As mentioned earlier, there are two assumptions that need to be considered by researchers and one of the assumptions is normality of the distribution. Unfortunately, the distributions of data in a real world are rarely normal. So, it is important to study the performance of the Parametric Bootstrap test in terms of the ability to control the Type I error where distribution is not normal.

Thus, in 2012, Parametric Bootstrap procedure with robust estimator namely trimmed mean and Winsorized variances was proposed by Cribbie *et al*. (2012). This procedure used Parametric Bootstrap test that recommended by Krishnamoorthy *et al.* (2007) as test statistic except that the means and variances were replaced by trimmed mean and Winsorized variances. They compared the modified Parametric Bootstrap test with the original Parametric Bootstrap test, original Welch test, Welch test with trimmed mean and James's second-order test. The results showed that Parametric Bootstrap test with trimmed mean and Winsorized variances provided a good control of Type I error and produced more powerful test than the original Parametric Bootstrap test when comparing the equality of means under the condition of non-normal distribution and unequal variances.

The study done by Cribbie *et al.* (2012) only focused on symmetric trimming where the proportion of data to be trimmed for each tail is the same which is 20%. There are some issues that need careful consideration when using symmetric trimming. First, symmetric trimming method used by Cribbie *et al.* (2012) becomes less efficient when the proportion of outliers in one tail of the distributions exceeds 20% especially for the extremely skewed distributions. It is because, firstly, the researcher

already fixed the amount of symmetric trimming percentage which is 20% for both of the tails of the distribution without looking at the shape of those distributions. Secondly, if the distribution is highly skewed to the right, it seems more reasonable to trim more observations from the right tail than trim equally from both tails of the distribution. In addition, how well Parametric Bootstrap test with symmetric trimmed mean can be performed in terms of Type I error compared to classical parametric test and non-parametric test were not explored by Cribbie *et al.* (2012).

Unlike the symmetric trimming, asymmetric trimming allows different number of observations that should be trimmed from each tail based on the characteristic of the data. However, the total number of trimmed data from the right and left tail must be equal to the total amount of trimming that is determined earlier. Asymmetric trimming strategy is similar with symmetric trimming where the number of observations that need to be trimmed for both methods still used predetermined trimming percentages. Thus, this method also cannot handle a situation where the outlier happens to exceed the predetermined amount of trimming.

Currently, there is new procedure that was developed to deal with the problem of predetermined amount of trimming which is modified one-step *M*-estimator (*MOM*) that was introduced by Wilcox and Keselman in 2003. This method was proved to be able to control of Type I error rates when testing the equality of central tendency under asymmetric distribution and variances heterogeneity (Syed Yahaya *et al*., 2006).

Therefore, this study proposed a modification of Parametric Bootstrap test introduced by Krishnamoorthy *et al.* (2007) with *MOM* which does not need a priori set the amount of trimming.

1.4 Modified One-Step *M*-estimator (*MOM*)

MOM approach is a technique to check the outliers in the data set, remove if outliers exist and average the remaining data. *MOM* allows for symmetric or asymmetric trimming, and has reputation for no trimming at all. Hence, *MOM* is flexible in handling outliers in a data with empirically determines the amount of trimming percentage regarding the shape of distribution.

1.5 Scale Estimators

The amount of trimming percentages for *MOM* estimator is empirically determined using robust scale estimator. In choosing the best scale estimator, the main factor to be considered is the value of a breakdown point. Several scale estimators such as MAD_n, S_n, Q_n, T_n and LMS_n have been introduced by Rousseeuw and Croux (1993). According to Rousseeuw and Croux (1993), MAD_n and T_n have the best possible breakdown value of 50%, and bounded influence function, with sharpest possible bound among all scale estimators.

Based on the study conducted by Syed Yahaya (2005), MAD_n and T_n were shown to have the ability in controlling the Type I error rates in testing the central tendency measure by using these scale estimators with S_1 statistics. Apart from that, the results from Md Yusof (2009) also indicated that by using MAD_n and T_n as robust scale estimators, the Type I error rates can be controlled insample with non-normal distribution and heterogeneous variances.

Motivated by the good performance of these procedures, a modification of the Parametric Bootstrap test statistic introduced by Krishnamoorthy *et al.* (2007) is proposed with *MOM* estimators, MAD_n and T_n as trimming criteria.

1.6 Problem Statement

The study done by Krishnamoorthy *et al.* (2007) did not explore the performance of the Parametric Bootstrap test under condition of non-normal distribution. Hence, in 2012, Cribbie *et al.* (2012) in their study proposed the Parametric Bootstrap test with 20% symmetric trimmed mean under non-normal distribution and variance heterogeneity. However, symmetric trimmed mean become less efficient when the proportion of outliers exceeded 20% and if the distribution is highly skewed to the right, it seems more reasonable to trim more observations from the right tail than trim equally from both tails of the distribution. In contrast to symmetric trimmed mean, asymmetric trimmed mean allows different number of observations that should be trimmed from each tail based on the characteristic of the data. However, asymmetric trimmed mean also cannot handle situation where the outliers happen to exceed the predetermined amount of trimming. Therefore, this study proposed a modification of Parametric Bootstrap test with *MOM* which does not need a priori set the amount of trimming and trimming based on the shape of distribution.

1.7 Objectives

The main objective for this research is to construct a robust test for independent groups as an alternative to the classical parametric test. In order to obtain this main objective, four sub-objectives are used. These four sub-objectives are as:

I. to modify the Parametric Bootstrap test by substituting the existing location estimator with *MOM* and the scale estimator with robust scale estimator, MAD_n or T_n.

II. to evaluate the performance of the modified Parametric Bootstrap test based on the Type I error rates by using simulated data.

III. to compare the performance of the modified Parametric Bootstrap test with the most frequently used classical parametric test (i.e. ANOVA and independent sample *t*-test) and non-parametric test (i.e. Mann Whitney test and Kruskal Wallis test) in terms of the Type I error rates.

IV. to investigate the performance of the modified Parametric Bootstrap test using real data.

1.8 Significance of the Study

This study can significantly contribute in experimental design. Normality and variance homogeneity are two main assumptions that need to be fulfilled when using classical parametric test such as independent sample *t*-test and ANOVA test. Experimental design methodology largely depends on it and not all real data is really encompassed with these two assumptions. In order to solve the problem, this study proposed a procedure that will not be constrained with all the assumptions. They can be used without the concern of normality of distribution and equality of variances.

Apart from that, with the modified procedure, researchers can increase accuracy in data analysis. While using classical parametric test, violation in any assumptions can falsely drawing conclusion based on the result obtained. Besides, violation in any assumptions can reduce the level of statistical power. Thus, the modified test

statistics are robust and can tolerate with all the assumptions and at the same time maintaining the accuracy of data analysis.

1.9 Organization of the Thesis

Chapter One gives the introduction of the study. This chapter briefly explained the importance of the study regarding the use of robust statistical analysis under the presence of non-normality and variances heterogeneity. This chapter also presents introduction on the proposed method namely Parametric Bootstrap test. In Chapter Two, a depth explanation on the proposed method is discussed. Chapter Two discusses about trimming and scale estimators used in this study. All the conditions that have been manipulated such as number of groups, balanced and unbalanced sample sizes, and type of distribution, variances heterogeneity and the nature of pairings are found in Chapter Three. Chapter Three presents the design specifications, explanation on the data generation and the proposed bootstrapping method used in this study. The Type I error rates for each procedures was presented in Chapter Four. Analysis on real data is recorded in Chapter Four. Lastly, we concluded our findings and proposed some recommendations for further studies in the last chapter which is Chapter Five.

CHAPTER TWO

LITERATURE REVIEW

2.1 Introduction

Classical parametric test such as independent sample *t*-test and ANOVA test are frequently used statistical method to test the equality of central tendency measures for two groups or more than two groups. However, the data obtained by researchers sometimes can be non-normal distribution and unequal variances (heterogeneous). Consequently, invalid results will be produced by the classical parametric test. It is because, violation in the assumption of normality and variances heterogeneity can distort the Type I error rates and will reduce the statistical power. The Type I error rates will be inflated from the nominal value when dealing with non-normal distribution and unequal variances and hence will falsely drawing conclusion. A lot of published articles have proved that violation in the assumption of classical parametric test can give biased results (Muhammad Di, Syed Yahaya & Abdullah, 2014; Md Yusof *et al.*, 2012a; Wilcox, 2002; Lix & Keselman, 1998; Micceri, 1989).

2.1.1 Non-normality

Skewness and outliers are two major problems that can make the distribution become non-normal. Wilcox and Keselman (2003) stated in their study that skewness can cause problems in controlling the Type I error. Skewed data will produce highly inaccurate Type I error rates and the confidence intervals. Skewness value for normal distribution and any symmetric distribution are zero. For asymmetric distribution, the value of skewness can be positive or negative. The value of skewness will be negative or less than zero when the data are skewed to the left. It means that the left tail is longer than the right tail. While positive value of skewness or more than zero

indicated that the data are skewed to the right meaning that the right tail for this distribution is longer than the left tail. Thus, when the data set are skewed, the rate of the Type I error and the statistical power will be affected. It is because the sample mean used in most statistical analysis is sensitive to the presence of outliers and skewness. Therefore, under non-normal distribution and heterogeneous variances, the test that used sample mean as estimator can become seriously inflated and will produce distorted Type I error rates with low statistical power.

The probability of the Type I error will be less than significant level when outliers exist on a data set and at the same time will reduce the power of the test statistic. According to Lix and Keselman (1998), the existence of extreme observations known as outliers on a distribution scores will influence the usual population standard deviation and hence will reduce the statistical power in detecting the differences between groups. Besides, the presence of outliers will lead to the observed scores being skewed. Wilcox and Keselman (2003) stated that simple transformation such as logarithm and square root failed to deal directly with outliers even though simple transformation can alter skewed distributions to become more symmetrical. Thus, they recommend using trimming method when dealing with outliers. It is well known that classical parametric test largely dependent on normality assumptions. The rate of the Type I error and power of the statistical test will be affected when this assumption is violated.

2.1.2 Heteroscedasticity

Another problem with classical parametric testoften encountered by researchers is heteroscedasticity. Heteroscedasticity can cause the classical parametric test to be biased even when the groups have a normal distribution (Wilcox & Keselman, 2003).

Box (1954) was found to be able to provide a good Type I error control while dealing with unequal variances. Unfortunately, this test is not suitable for extreme variances heterogeneity. It is because Box's numerical results were based on the situations where the ratio of the largest standard deviation of the groups that being compared with the smallest variances did not exceed $\sqrt{3}$. This method has difficulty in controlling the Type I error rates when the ratio increase more than $\sqrt{3}$ and the group distributions are non-normal. Some researcher proposed to use non-parametric test as alternative such as Mann Whitney and Kruskal Wallis. However, these methods have low statistical power (Syed Yahaya, 2005). Apart from that, a lot of published articles showed that non-parametric test are not robust when dealing with heteroscedasticity data (Lix *et al.,* 1996 and Zimmerman, 2000).

For this reason, many researchers had contributed in the development of alternative approach with robust procedures to deal with the problems of non-normality and heteroscedasticity. According to Hampel (2001), robust statistics is the stability theory of statistical procedures. It means that the statistical procedures insensitive to the violation of non-normality and unequal variances and hence will provide a good Type I error control. Wilcox (2005) in their study stated that the term robust statistics refers to procedures that are able to maintain the Type I error at its nominal level and at the same able to maintain the statistical power even under the condition of non-normal distribution and variances heterogeneity.

In this study, Parametric Bootstrap procedure is proposed as a test statistic to handle the problem of non-normality and variances heterogeneity. Krishnamoorthy *et al.* (2007) proposed Parametric Bootstrap test as a new test statistic to test the equality of central tendency measures that is mean under heterogeneous variances in the

groups. Based on the results obtained, Parametric Bootstrap test intended to provide a good control of the Type I error rates and more powerful than the other tests namely original Welch test, James test and the generalized F (FG) test.

2.2 Trimming

Regarding the weaknesses of non-parametric test and simple transformation explained earlier in **CHAPTER 1**, trimming is an alternative to reduce the impact of outliers. This trimming method can eliminate outliers or extreme observations in a data set. The correct choices of the amount of trimming can be very beneficial in terms of efficiency (Keselman, Kowalchuk, Algina, Lix & Wilcox, 2000). By efficiency, it means achieving a relatively small standard error. Keselman *et al.* (2000) concluded that efficiency can be poor when sampling from a heavy-tailed distribution with too small trimming percentages. However, if trimming percentages is too large, the efficiency will go down when sampling from a normal distribution. More trimming is beneficial if there are large number of outliers tend to appear (Wilcox, 2010). Hence, the problem of non-normality can be reduced when more trimming is applied. However, Cribbie *et al.* (2012) has expressed concern that too much trimming will reduce power while too little trimming would not provide a good control of Type I error.

There are two approaches in trimming which are symmetric trimming and asymmetric trimming. Symmetric trimming trims the same amount of observations from the right tail and the left tail of the distribution. The amount of symmetric trimming percentage was determined earlier by researchers. This method was simple and very convenience for data analyzing. Symmetric trimming is quite efficient for symmetric distribution because it trims the same amount of observations at both ends

of data. Wilcox (2003) stated that, when sampling from symmetric distribution, it is intuitively appealing to use symmetric trimming.

There are different trimming percentages suggested by different researchers. For example, Babu, Padmanabhan and Puri (1999) suggested 15% trimmed mean in order to obtain a good control of the Type I error. However, Wilcox (2003) and Rosenberger and Gasko (1983) in their study found that 20% trimmed mean can show a good Type I error control and statistical power under the conditions of non-normal distributions and unequal variances. Rocke, Downs and Rocke (1982) recommended the used of 20% - 25% symmetric trimming. Based on the literature, there are many trimming percentages that can be adopted by researcher and not constrained to only one percentage of trimming.

However, symmetric trimming become less efficient when there is even just a slight departure from symmetric such as only containing one outlier value (Wu & Zuo, 2009). In addition, according to Keselman *et al.* (2002), there are two practical concerns that need to be considered when using symmetric trimming as detailed below;

i. the proportion of outliers can exceeds the percentage of symmetric trimmed mean and hence will require more than the amount of trimming that determined earlier. For example, when the trimming percentage is set at 20%, more amount of trimming percentage is needed if the outliers exceeded 20% from both tail of distribution.
ii. the distribution can either be negatively or positively skewed. When the distribution is negatively skewed, more amounts of observations should be trimmed from the left tail compared to the right tail of the distribution. On the

other hand, more amounts of observations should be trimmed from the right tail compared to the left tail if the distribution is skewed to the right.

Another approach in trimming is asymmetric trimming. Unlike symmetric trimming, asymmetric trimming allows for different amount of trimming percentage for each tail of distribution. Thus, asymmetric trimming has been theorized to be potentially advantageous when the distributions are known to be skewed since not all of the data are symmetric (Micceri, 1989). That means extremely skewed distribution needs to be trimmed more than normal distributions because extremely skewed distribution contains more outliers or extreme values compared to normal distribution.

Previous researchers have identified asymmetric trimming as trimming method that may provide a successful solution in controlling the Type I error under the presence of non-normal distributions and unequal variance in the groups (Babu *et al.*, 1999). However the amount of trimming is predetermined for each tail of distribution before the trimming process is performed. In other words, the total number of trimming from the left and right tail of the distribution must be equal to the amount of trimming that was determined earlier. This method cannot handle the situation where the outlier happens to exceed the predetermined amount of trimming for both tail of distributions.

Therefore, modified one-step *M*-estimator (*MOM*) can be used to avoid the problem of predetermined amount of trimming. *MOM* was recommended by Wilcox and Keselman (2003) as new trimming strategy that does not need a priori set of trimming percentage. Thus, *MOM* can give more advantages compared to predetermined trimming. Besides, *MOM* is a trimming strategy that trimming based

on the shape of the distributions. Therefore, in this study we will use the Parametric Bootstrap test with several *MOM* estimators namely MAD_n and T_n as trimming criteria proposed by Rousseeuw and Croux (1993). Then, the Type I error rates for these procedures are examined and compared to the classical parametric test and non-parametric test.

Before going deeply into the discussion of the proposed test statistic with the selected robust scale estimators, few terminologies of this study is briefly discussed in the next section.

2.3 Type I Error

Type I error can be defined as the probability of rejecting null hypothesis even though it is true. The Type I error occurs when the decision to reject the null hypothesis is incorrect. Null hypothesis is a statement about population parameter that always assumed to be true. The population parameter that always been used is mean, median and variances. The Type I error rate is designed by the Greek letter alpha (α).

Conventionally, if the Type I error rates produced by a procedure fall between 0.5α and 1.5α, the procedure can be considered robust (Bradley, 1978). In this study, the significance level is set at $\alpha = 0.05$. Therefore, the Type I error rate should be in between 0.025 and 0.075. Mehta and Srinivasan (1970) in their study stated that a procedures still could be considered robust if the true Type I error rate is equal or less than the significance level. In addition, according to Guo and Luh (2000), a procedure with the empirical Type I error below the 0.075 level can be considered robust if the significant level is 0.05.

2.4 Breakdown Point

One of the most popular measures of robustness of a statistical procedure is the breakdown point. Wilcox (1997) stated that the breakdown point refers to the smallest proportion of observations, that when altered sufficiently, can render the estimator meaningless. The sample value that uncharacteristically large or uncharacteristically small will make the estimator break down.

The breakdown point of the sample mean $\overline{X}$ is only $1/n$ because only one data point from n observations needed to be replaced to force the sample mean arbitrarily further from the true mean. As the ith observation among the observations $X_1,\ldots,X_n$ increases to infinity, the sample mean increases to infinity as well and the breakdown point of the sample mean equal to zero, because $1/n$ tends to 0. It means that even single outlier can break it down hence the usual sample mean is not robust. In contrast to sample mean, the breakdown point of the γ trimmed mean is γ. For example, the breakdown point for 10% trimmed mean is 0.10. Thus, the trimmed mean will be moved away from the true mean when more than 10% data points are altered.

The estimators can withstand large proportions of very bad observations without breaking down completely when the value of breakdown point is high. The sample median has a breakdown point of 0.5, which is the highest possible value. It means that more than 50% data points have to be replaced with values arbitrarily far from zero in order to move the sample median arbitrarily further from the true median.

2.5 Influence Function

Another property for judging robustness is the influence function. Influence function is the derivative of a functional T(F) introduced by Hampel (1974). The influence function measures the relative extent a small perturbation in F has on T(F) (Staudte & Sheather, 1990). The influence function measures the change in the functional due to a small amount of contamination at the point *x*. Attention might be restricted to those measure having a bounded influence function if the goal is to minimize the influence of a relatively small number of observations on a measure of central tendency. Thus, a T(F) with an influence function that is bounded in *x* is more robust to extreme value. A robust estimator means that the influence of any single observation is insufficient to yield any significant offset.

2.6 Central Tendency Measures

Measures of central tendency are measures of central location of a distribution. A measure of central tendency refers to a single value or middle value of a data set. This value used to describe the data set. The most commonly used measures of central tendency are mean, mode and median. Different calculation is needed for different measure of central tendency. Under certain situation, some measures of central tendency can perform better than others.

The mean is the most familiar and well known compared to others measures of central tendency because the fact that it makes use of all the values in a distribution (Miller & Brewer, 2003). The mean can be calculated by dividing the sum of all the values in the data set divided with the *n* value in that set of data. However, this measure of central tendency is very sensitive to extreme values. Based on its breakdown point which is zero, only one single outlier is needed in order to move the

sample mean far from the actual mean. In addition, this estimator has unbounded influence function meaning that a single contaminated observations may have a considerable effect on the estimate. Therefore, any methods that based on the sample mean will produce low power and the rates of the Type I error can be distorted. For this reason, Erceg-Hurn and Mirosevich (2008) recommended to the use of modern robust statistical method instead of classical parametric test by using a wide range of software. For example, Cribbie *et al.* (2012) suggested a robust approach such as trimmed mean as central tendency measure to hypothesis testing.

The central tendency measure used in this study is *MOM*. By replacing the usual mean with *MOM*, tests that are insensitive to both non-normality and variances heterogeneity can be obtained.

2.6.1 Modified one step *M*-estimator (*MOM*)

MOM is a strategy to check the outliers in a data set. Later remove the outlier whether to trim symmetrically or asymmetrically and average the remaining values. *MOM* was introduced by Wilcox and Keselman (2003) by modifying one-step *M*-estimator. The one-step *M*-estimator can be defined as

$$\bar{X}_{tj} = \frac{1.28(MAD_{n_j})(i_1 - i_2) + \sum_{i=i_i+1}^{n_j - i_2} X_{(i)j}}{n_j - i_1 - i_2}$$

Wilcox and Keselman (2003) in their study found that the one-step *M*-estimator perform reasonable well in terms of Type I error only when the sample sizes more than 20 ($n > 20$). This method failed to perform well for the sample sizes less than 20. Thus, they modified the one-step *M*-estimator by dropping the term containing

MAD_n. *MOM* showed a good control on Type I error and statistical power even with small sample sizes.

The *MOM* estimator recommended by Wilcox and Keselman (2003) can be defined as:

$$\bar{X}_{tj} = \frac{1}{n_j - i_1 - i_2}\left[\sum_{i=i_1+1}^{n_j-i_2} X_{(i)j}\right]$$

where

$X_{(i)j}$ is the i^{th} ordered observations in group j

$\widehat{M}_j$ is median of group j

n_j is the number of observation for group j

i_1 is the number of extreme observations on the left tail such that

($X_{(i)j} - \widehat{M}_j$) < -K(scale estimator),

i_2 is the number of extreme observations on the right tail such that

($X_{(i)j} - \widehat{M}_j$) >K(scale estimator)

In the case of one-step *M*-estimator, the *K* value is 1.28 which is the 0.9 quantile of the standard normal distribution (Wilcox, 1997). Then, Wilcox and Keselman (2003) adjusted the *K* value so that efficiency is good under normality especially for small sample sizes. Using simulations with 10,000 replications, they found that the standard error of the sample mean divided by the standard error of $\bar{X}_{tj}$ is approximately 0.9 for $n_1 = n_2 = n_3 = n_4 = n_5 = 20$ with $K = 2.24$. Therefore, they suggested the scale estimator multiply with the value of 2.24 for trimming criteria. For *MOM*, $K = 2.24$ is constant in order for having a reasonably small standard error

when sampling from a normal distribution (Othman, Keselman, Padmanabhan, Wilcox & Fradette, 2004). For a sample with no extreme value, i_1 and i_2 is equal to zero where *MOM* is equal to the mean for the group.

2.7 Scale Measures

Measures of central tendency do not describe the whole picture of distribution. It is because data which have the same average could have very different spreads. A scale measure is the measures that acknowledge the spread of the data. The term dispersion indicates the spread or variation in the values of a variable (Miller & Brewer, 2003). The range, the variance and the standard deviation are commonly used as measure of dispersion. However, like the range, the variance and standard deviation can be affected by the presence of outliers or extreme values in data set. Therefore, these scale measures are not robust. In this study, the Winsorized variances is adopted in order to get tests that are insensitive to the effects of non-normality and variance heterogeneity.

2.7.1 Winsorized Variances

Trimming and winsorization are methods for reducing the effect of outliers in sample data. For example, Cribbie *et al.* (2012) found that under non-normal distribution and heterogeneous variances, a statistic with trimmed mean and Winsorized variances could perform better in controlling the Type I error rates and the statistical power compared to statistics based on the usual mean and variance.

The trimmed mean must be calculated first in order to get Winsorized variances. The Winsorized variances is computed after the smallest non-trimmed score replaced the removed scores from lower tail of distribution, and the highest non-trimmed score

replaced the removed scores from the upper tail of distribution. The non-trimmed and replaced scores from both tails of distribution are called Winsorized score. Then, in order to calculate the Winsorized mean, the Winsorized scores will be divided by n_j. Lastly, the the sum of squared deviations of Winsorized scores from the Winsorized mean will be divided by $n_j - 1$. This value is called Winsorized variance. The value of Winsorized variances can be computed by

$$\frac{\sum_i (Y_i - \bar{X}_w)^2}{n_j - 1}$$

2.8 Robust Scale Estimators

To choose a robust scale estimator, there are two factors that should be considered. First is the value of a breakdown point (Wilcox, 2005). Several robust scale estimators with highest breakdown point have been introduced by Rousseeuw and Croux (1993) such as, MAD_n, S_n, Q_n, T_n and LMS_n. By referring to breakdown point, MAD_n and T_n have the best possible breakdown point that is 50%, twice as much as interquartile range. Second factor that need to be considered while choosing these scale estimators are their bounded influence function. MAD_n and T_n also exhibit bounded influence function. Based on these two factors and their good performance in Syed Yahaya (2005), MAD_n and T_n are chosen for this study.

2.8.1 MAD_n

MAD_n is the median absolute deviation about the median which is a very popular robust scale estimator. This scale estimator demonstrated the best possible breakdown point (0.5) and its influence function is bounded with the sharpest bound among all scale estimators. According to Huber (1981), MAD_n is a single most useful ancillary estimate of scale. Furthermore, MAD_n is simple, easy to compute and very

useful. Their extreme sturdiness makes MAD_n ideal for screening the data for outliers in a quick way by computing

$$\frac{|x_i - med_j x_j|}{MAD_n} > K$$

where this robust scale estimator is given by

$$MAD_n = b \text{ med}_i | x_i - med_j x_j |$$

The constant b = 1.4826 is needed to make the estimator consistent for the parameter of interest, $x_i = x_1, x_2, \ldots, x_n$ and $i > j$. However, MAD_n is not without disadvantage. Its efficiency at Gaussian distribution is very low with only 37% efficient.

2.8.2 T_n

T_n is another promising scale estimator proposed by Rousseeuw and Croux (1993). With 52% efficiency at Gaussion, T_n can be considered more efficient than MAD_n.

Given as

$$T_n = 13800 \frac{1}{h} \sum_{k=1}^{h} \{ med_{j \neq 1} | x_i - x_j \}_{(k)}$$

where

$$h = \left[\frac{n}{2}\right] + 1$$

Apart from that T_n is a scale estimator that demonstrated highest breakdown point like MAD_n and a bounded influence function. Compared to other scale estimators, the calculation of T_n is much easier and it is suitable for asymmetric distributions.

2.9 Statistical Methods

This study focuses on the Parametric Bootstrap statistic introduced by Krishnamoorthy *et al.* (2007). Krishnamoorthy *et al.* (2007) found that this test statistic was able to control Type I error rates when conducting test on the effect of variances heterogeneity. For this study, the Type I error rates of this test are examined under conditions of homogeneous and heterogeneous variances across three types of distribution (i.e. normal, moderately skewed and extremely skewed). Thereafter, this test was compared with parametric and non-parametric test in terms of Type I error rates in order to determine the best procedure. The T_{N0} statistics proposed by Krishnamoorthy *et al.* (2007) is given by

$$T_{N0} = \sum_{i=1}^{k} \frac{n_i}{s_{Bi}^2} \bar{X}_{Bi}^2 - \frac{\left(\sum_{i=1}^{k} (n_i \bar{X}_{Bi} / s_{Bi}^2)\right)^2}{\sum_{i=1}^{k} n_i / s_{Bi}^2}$$

In this study, the bootstrap method which is percentile bootstrap is used to obtain the *p*-values which is represented by (number of $T_{NOB}^{*} > T_{NO}$) / B (discussed briefly in CHAPTER THREE).

2.10 Bootstrapping

Efron (1979) introduced the bootstrap method that can be used as a computer-based method for estimating the standard error of $\widehat{\theta}$. This method spread widely in empirical research. According to Staudte and Sheather (1990), the word bootstrap is used to indicate that the observed data are used to obtain an estimate of the parameter and to generate new samples.

A pseudo sampling distribution of the estimator can be estimated using bootstrap when the sampling distribution of the estimator is unknown. Variability of an

estimator, bias of an estimator and significance of a test involving the estimator can be assessed with the establishment of the pseudo sampling distribution.

Apart from that, according to Babu *et al*. (1999), the bootstrap method is known to give a better approximation than one based on the normal approximation theory. Bootstrap method can improve the ability of the test in controlling the Type I error compared to non-bootstrap methods (Othman, Keselman, Padmanabhan, Wilcox & Fradette, 2003). Results of Westfall and Young (1993) suggested that the combination of bootstrap methods with methods based on trimmed means could improves the Type I error control. Further improvement in Type I error control is often possible by obtaining critical values for test statistic through bootstrap (Keselman, Wilcox & Lix, 2003). The bootstrap procedures on Parametric Bootstrap test are discussed briefly in CHAPTER THREE.

CHAPTER THREE

METHODOLOGY

3.1 Introduction

Classical parametric test such as independent sample *t*-test and ANOVA test often used to test the equality of central tendency measures for independent groups. However, violation in the assumptions of normality and equality of variances can distort the Type I error rates. Therefore, our main focus for this study is to use robust Parametric Bootstrap test to test the equality of means for independent groups that can tolerate the violation of these assumptions. Parametric Bootstrap test was used with robust scale estimators namely MAD_n and T_n as trimming criteria to trim data empirically. This test statistics use group trimmed means as the central tendency measures.

3.2 Procedures Employed

This study modified Parametric Bootstrap test using *MOM* estimator. This trimming strategy involved robust scale estimator namely MAD_n and T_n. The modified Parametric Bootstrap test with its corresponding scale estimators is shown in *Figure 3.1*.

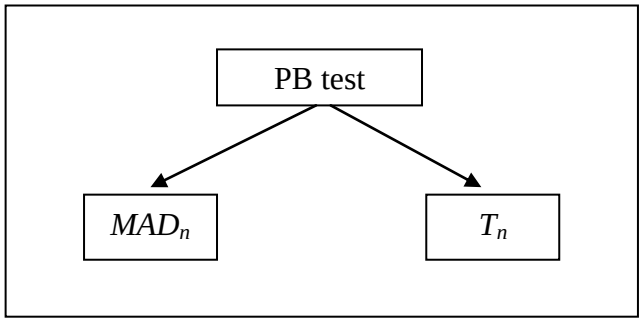

Figure 3.1. Statistical test with the corresponding scale estimators

The performance of the combination of the Parametric Bootstrap test with the aforementioned scale estimators namely MAD_n and T_n were investigated and compared with classical parametric and non-parametric tests in terms of Type I error rates.

3.2.1 Parametric Bootstrap Test with MAD_n

The Parametric Bootstrap test is an alternative test that was shown to provide a good Type I error control even for small sample sizes and large number of groups under the condition of unequal variances (Krishnamoorthy *et al.*, 2007). In order to make Parametric Bootstrap test robust to the conditions of non-normal distribution and unequal variances, the usual means will be replaced by *MOM* estimators and the usual variances will be replaced by the modified Winsorized variances. Let $X_{(1)j}$, $X_{(2)j}$,…, $X_{(nj)j}$ represent the ordered sample of group j with size n_j.

The first step in the modified Parametric Bootstrap procedure with MAD_n is the calculation of *MOM* of group j by using:

$$\bar{X}_{tj} = \frac{1}{n_j - i_1 - i_2}\left[\sum_{i=\, i_i+1}^{n_j - i_2} X_{(i)j}\right] \qquad [3.1]$$

where

i_1 is the number of observations $X_{(i)j}$ such that

$(X_{(i)j} - \widehat{M}_j) < -2.24(MAD_n)$,

i_2 is the number of observations $X_{(i)j}$ such that

$(X_{(i)j} - \widehat{M}_j) > 2.24(MAD_n)$

$\widehat{M}_j$ is median of group j

n_j is group sample sizes

$(MAD_n)_j$ is median absolute deviation about the median of group j

The median absolute deviation (MAD_n) is simple and easy to compute given by:

$$MAD_n = b \text{ med}_i | x_i - med_j x_j | \quad [3.2]$$

The constant *b* in the equation above is needed to make the estimator consistent with the parameter of interest.

Then, the usual variances will be replaced by the modified Winsorized variances where the equation is given by:

$$s_{t_j}^2 = \frac{s_{W_j}^2(n_j)}{h_j - 1} \quad [3.3]$$

where the sample Winsorized variance is

$$s_{W_j}^2 = \frac{\sum_i (Y_i - \bar{X}_w)^2}{n_j - 1}$$

and

$$\bar{X}_W = \frac{1}{n}\sum_i Y_i \quad [3.4]$$

Lastly, compute T_{N0} statistics (Krishnamoorthy*et al.*, 2007) of group *j* by using

$$T_{N0} = \sum \frac{n_j}{s_{t_j}^2} \bar{X}_j^2 - \frac{\left(\sum_j (n_j \bar{X}_j / s_{t_j}^2)\right)^2}{\sum_j n_j / s_{t_j}^2} \quad [3.5]$$

3.2.2 Parametric Bootstrap Test with T_n

The first step in the modified Parametric Bootstrap procedure with T_n is the calculation of *MOM* of group *j* by using:

$$\overline{X}_{tj} = \frac{1}{n_j - i_1 - i_2}\left[\sum_{i=\, i_i+1}^{n_j - i_2} X_{(i)j}\right] \quad [3.6]$$

where

i_1 is the number of observations $X_{(i)j}$ such that

($X_{(i)j} - \widehat{M}_j$) < -2.24(T_n),

i_2 is the number of observations $X_{(i)j}$ such that

($X_{(i)j} - \widehat{M}_j$) > 2.24(T_n)

$\widehat{M}_j$ is median of group *j*

n_j is group sample sizes

$(T_n)_j$ is robust scale estimator of group *j*

The calculation of T_n is given by

$$T_n = 13800\,\frac{1}{h}\sum_{k=1}^{h}\left\{med|x_i - x_j\right\}_{(k)} \quad [3.7]$$
$$\text{j} \neq 1$$

where

$$h = \left[\frac{n}{2}\right] + 1$$

After the calculation of *MOM* of group *j*, we proceed with the computation of T_{NO} given in the equation [3.5].

3.3 Variable Manipulated

Each procedure has been investigated based on the conditions resulted from manipulation of five variables such as number of groups, group sample sizes (i.e. balance and unbalanced), type of distribution (i.e. normal and non-normal), variances (i.e. homogeneity and heterogeneity) and nature of pairings (i.e. positive and negative) in order to highlight the strengths and weaknesses of the test in testing the

equality of central tendency measures. This manipulation helps to identify the robustness of the proposed test when dealing with the problems of non-normality and heterogeneity. All of the outcomes from these different conditions were compared in terms of the Type I error rates. Table 3.1 represents the summary of all the conditions used in this study.

Table 3.1

Description of Variable Manipulated

Conditions	Descriptions
Number of Groups	$J = 2$
	$J = 3$
Sample Sizes	Balanced
	Unbalanced
Type of Population Distributions	Normal
	Moderately skewed
	Extremely skewed
Variances	Equal
	Moderate
	Large
Nature of pairings	Positive
	Negative

3.3.1 Numbers of Groups

The number of groups containing randomized design of two groups ($J = 2$) and three groups ($J = 3$). The difference in number of groups is to represent study with the procedure of two groups and more than two groups. Investigation on two groups ($J = 2$) is chosen because there are a lot of previous work related to this study such as by Md Yusof, Othman and Syed Yahaya (2010), Lix and Keselman (1998), Md Yusof

et al. (2012a) and Md Yusof, Abdullah and Syed Yahaya (2012b) had also utilized similar design.

Apart from that, investigation on three groups ($J = 3$) is chosen since the previous research done by Cribbie *et al*. (2012) showed that Parametric Bootstrap test with trimmed mean able to produce Type I error rates close to nominal value of $\alpha = 0.05$ by using this number of groups. By analyzing on the different number of groups, we are able to examine the effect of the number of groups on the Type I error rates for each procedure.

3.3.2 Balanced and Unbalanced Sample Sizes

Group sample sizes were set balanced and unbalanced for the purpose of examining the effect of sample sizes on the Type I error rates for the case of ($J = 2$) and ($J = 3$). For balanced sample sizes, total sample sizes for $J = 2$ was set at 40. The number of observation for each group for the case of ($J = 2$) is the same which is $n_1 = n_2 = 20$. A study done by Syed Yahaya (2005) showed that this number of sample sizes provided a good control of Type I errors. For the case of three groups ($J = 3$), the total sample sizes was set at 60 where the number of observation for each group was set to be equal to 20 ($n_1 = 20$, $n_2 = 20$, $n_3 = 20$). The number for each group for the case of ($J = 3$) was set to be the same with ($J = 2$) case.

In contrast to balanced sample sizes, the number of observation or unbalanced sample sizes was set at different sample sizes for each group which is $n_1 = 15$ and $n_2 = 25$ for the case of two group ($J = 2$). While for three groups ($J = 3$), the number of observation for each group was set at $n_1 = 15$, $n_2 = 20$, $n_3 = 25$ with subsequent

increment of 5 for each group. Table 3.2 summarizes the design of each number of groups for balanced and unbalanced sample sizes.

Table 3.2

Balanced and Unbalanced Sample Sizes

***J* = 2**		***J* = 3**	
Balanced	**Unbalanced**	**Balanced**	**Unbalanced**
n_1= 20	n_1= 15	n_1= 20	n_1= 15
n_2= 20	n_2= 25	n_2= 20	n_2= 20
		n_3= 20	n_3= 25
Total = 40	Total = 40	Total = 60	Total = 60

3.3.3 Types of Distributions

Another condition that is considered in this study is types of distributions. As mentioned earlier in CHAPTER ONE, normal distribution is one of the two major assumptions that need to be satisfied before proceeding with the ANOVA test and independent sample *t*-test. However, the assumption of normal distribution in a data set is rarely met. A slight departure from normal distribution can distort the Type I error rates and hence can lead to wrong conclusions. Thus, this study investigates the performance of the proposed procedure in terms of the ability to control the Type I error rates under various types of distribution. Three types of distribution were considered to represent different levels of skewness. The three types of distribution are:

i. Normal distribution

ii. Moderately skewed distribution

iii. Extremely skewed distribution

Introduced by Hoaglin (1985), the *g*- and *h*- distribution was used to represent the types of distributions. Cribbie *et al.* (2012), Md Yusof (2009) and Othman *et al.* (2004) used data generated from the *g*- and *h*- distribution in order to examine the effect of distributional shapes on Type I error. The parameter *g*- controls the value of skewness while *h*- controls the value of kurtosis or the amount of elongation. That is as *g*- increases, the distribution becomes increasingly positively skewed and the tails of distribution will become heavier as *h*- increases.

In this study, $g = 0.0$ and $h = 0.0$ for normal distribution. The zero value for *g* and *h* give the meaning that the distribution is symmetric and the tails are normally distributed. For the second distribution, $g = 0.5$ and $h = 0.0$ which represents moderately skewed distribution. It is because the tails of distribution become positively skewed as *g* increase. The skewness and kurtosis values for this distribution are $\gamma_1 = 1.74$ and $\gamma_2 = 8.9$; respectively (Othman *et al.*, 2004). For extremely skewed distribution, Cribbie *et al.* (2012) used $g = 1.0$ and $h = 0.0$ to represent the extremely skewed distribution with skewness of $\gamma_1 = 6.18$ and kurtosis of $\gamma_2 = 113.94$. Table 3.3 summarizes the information on *g*- and *h*-distribution which is used in this study.

Table 3.3

Summary of g- and h- Distribution

Groups	Distribution Shapes
***J* = 2**	$g = 0.0, h = 0.0$
	$g = 0.5, h = 0.0$
	$g = 1.0, h = 0.0$
***J* = 3**	$g = 0.0, h = 0.0$
	$g = 0.5, h = 0.0$
	$g = 1.0, h = 0.0$

3.3.4 Variances Heterogeneity

Variances heterogeneity is one of the two major assumptions that often violated when testing the equality of central tendency. The classical parametric test has been known to yield misleading results when the population variances differ. Hence, to investigate the effect of variances heterogeneity on Type I error rates, three natures of variances (i.e. equal variances, moderately unequal variances and largely unequal variances) were assigned to the groups.

In this study, variance with a ratio 1:36 (1:1:36) is used to represent largely unequal variances. According to Keselman, Wilcox, Algina, Fradette and Othman (2004), although the selected ratio may large, ratios similar to this case and larger have been reported in the literature. Keselman *et al.* (1998) also found a ratio of 24:1 and 29:1 after reviewing articles published in prominent education and psychology journals. Apart from that, Wilcox (2003) cited data sets where the ratio as high as 17,977:1. Therefore, it seems reasonable to investigate the robustness of each procedure under a potentially extreme condition even though the ratio of 1:36 (1:1:36) may be large. A procedure is likely to work under most conditions of heterogeneity which are

likely to be encountered by researchers if it works under an extreme degree of heterogeneity.

For moderately unequal variances, a ratio of 1:8 (1:8:16) was assigned to this study. Keselman, Wilcox, Lix, Algina, Fradette (2007) categorized 8:1 ratio as less extreme condition of heterogeneity. It is important for the comparative study to check performance of all procedures under any degree of heterogeneity which are likely to be encountered by researchers.

3.3.5 Nature of Pairings

Sample sizes with unequal variances can have two types of pairing, such as, positive pairing and negative pairing. For the case of balanced sample sizes, positive pairing resulted from the association of the balanced observations with the lowest and the highest variances while negative pairing resulted when the group having the balanced observations associated with the highest and the lowest variances. For the case of unbalanced sample sizes, for positive pairing, the group having the smallest observation will associate with the smallest variance and the largest observation will associate with the largest variance. In contrast, for negative pairing, the group having the smallest observation will associate with the largest variances while the largest observation will associated with the smallest variances. The nature of pairings does have potential to produce conservative and liberal results, respectively (Lix & Keselman, 1998; Othman *et al*, 2004). Therefore, in this study, we analyzed the robustness of each investigated procedure under the two types of pairings. Table 3.4 - Table 3.7 represent the pairing of variances and sample sizes that will be used in this study.

Table 3.4

Balanced Sample Sizes and Pairing of Variances for J = 2

Pairing	Group Sizes		Variance					
			Equal		Moderately Unequal		Largely Unequal	
	1	2	1	2	1	2	1	2
Positive	20	20	1	1	1	8	1	36
Negative	20	20	1	1	8	1	36	1

Table 3.5

Unbalanced Sample Sizes and Pairing of Variances for J =2

Pairing	Group Sizes		Variance					
			Equal		Moderately Unequal		Largely Unequal	
	1	2	1	2	1	2	1	2
Positive	15	25	1	1	1	8	1	36
Negative	15	25	1	1	8	1	36	1

Table 3.6

Balanced Sample Sizes and Pairing of Variances for J =3

Pairing	Group Sizes			Variance								
				Equal			Moderately Unequal			Largely Unequal		
	1	2	3	1	2	3	1	2	3	1	2	3
Positive	20	20	20	1	1	1	1	8	16	1	1	36
Negative	20	20	20	1	1	1	16	8	1	36	1	1

Table 3.7

Unbalanced Sample Sizes and Pairing of Variances for J = 3

Pairing	Group Sizes			Variance								
				Equal			Moderately Unequal			Largely Unequal		
	1	2	3	1	2	3	1	2	3	1	2	3
Positive	15	20	25	1	1	1	1	8	16	1	1	36
Negative	15	20	25	1	1	1	16	8	1	36	1	1

For both balanced and unbalanced sample sizes, in the case of $J = 2$, the variances for positive pairing were set at 1:8 ratio (moderately unequal) and 1:36 ratio (largely unequal). For negative pairing, the variances were set at 8:1 ratio (moderately unequal) and 36:1 ratio (largely unequal). For the case of $J = 3$, the variances for positive pairing were set at 1:8:16 ratio (moderately unequal) and 1:1:36 ratio (largely unequal) while for negative pairing, the variances were set at 16:8:1 ratio (moderately unequal) and 36:1:1 ratio (largely unequal).

3.4 Design Specification

Table 3.8

Design specification for balanced sample sizes and homogeneous variances for J = 2

$N = 40$			
Group sizes		*Group Variances*	
Group 1	Group 2	Group 1	Group 2
20	20	1	1

Table 3.9

Design specification for balanced sample sizes and heterogeneous variances for $J = 2$

$N = 40$				
Pairing	*Group sizes*		*Group Variances*	
	Group 1	Group 2	Group 1	Group 2
P	20	20	1	8
N	20	20	8	1
P	20	20	1	36
N	20	20	36	1

Table 3.10

Design specification for unbalanced sample sizes and homogeneous variances for $J = 2$

$N = 40$			
Group sizes		*Group Variances*	
Group 1	Group 2	Group 1	Group 2
15	25	1	1

Table 3.11

Design specification for unbalanced sample sizes and heterogeneous variances for $J = 2$

$N = 40$				
Pairing	*Group sizes*		*Group Variances*	
	Group 1	Group 2	Group 1	Group 2
P	15	25	1	8
N	15	25	8	1
P	15	25	1	36
N	15	25	36	1

Table 3.12

Design specification for balanced sample sizes and homogeneous variances for J = 3

N = 60					
Group sizes			*Group Variances*		
Group 1	Group 2	Group 3	Group 1	Group 2	Group 3
20	20	20	1	1	1

Table 3.13

Design specification for balanced sample sizes and heterogeneous variances for J = 3

N = 60						
Pairing	*Group sizes*			*Group Variances*		
	Group 1	Group 2	Group 3	Group 1	Group 2	Group 3
P	20	20	20	1	8	16
N	20	20	20	16	8	1
P	20	20	20	1	1	36
N	20	20	20	36	1	1

Table 3.14

Design specification for unbalanced sample sizes and homogeneous variances for J = 3

N = 60					
Group sizes			*Group Variances*		
Group 1	Group 2	Group 3	Group 1	Group 2	Group 3
15	20	25	1	1	1

Table 3.15

Design specification for unbalanced sample sizes and heterogeneous variances for J = 3

$N = 60$						
Pairing	*Group sizes*			*Group Variances*		
	Group 1	Group 2	Group 3	Group 1	Group 2	Group 3
P	15	20	25	1	8	16
N	15	20	25	16	8	1
P	15	20	25	1	1	36
N	15	20	25	36	1	1

3.5 Data Generation

This study will be based on simulation of the data that follow the conditions mentioned earlier. For data generation, this study will use SAS/IML version 9.3. According to Wilcox and Keselman (2010), skewness in a data can cause problem when trying to control the Type I error rates. Thus, to examine the effect of distributional shape on Type I error rates, the simulation data will follow the type of distributions chosen. In order to represent all types of distributions, *g*- and *h*-distribution is considered.

The following steps are used to generate the pseudo-random variates for the *g*- and *h*-distribution:

i. Generate standard normal variates (Z_{ij}) using SAS generator RANNOR (SAS Institute, 1999).

ii. Convert the standard normal variates to random variables via equation

$$X_{ij} = \begin{cases} \frac{\exp(gZ_{ij})-1}{g}\exp\left(\frac{hZ_{ij}^2}{2}\right) & g \neq 0 \\ Z_{ij}\exp\left(\frac{hZ_{ij}^2}{2}\right) & g = 0 \end{cases} \quad [3.8]$$

For the conditions of $g \neq 0$, the central tendency measure such as trimmed mean produced by *MOM* estimator is unequal to zero. Thus, the observations Y_{ij}, from each simulated distributions were standardized by subtracting the population central tendency parameter (ω) from each observation in order to make certain that the null hypothesis remains true such that,

$$Y_{ij} = X_{ij} - \omega \quad [3.9]$$

For robust scale estimator MAD_n, the values of the population trimmed mean were based on 1,000,000 observations. However, due to time constraint and the limited power of computer, the values of the population trimmed mean for robust scale estimator T_n, were generated from 100,000 observations only. A summary of the population trimmed mean corresponding to scale estimators for each type of distribution is presented in Table 3.16.

Table 3.16

Population trimmed mean for g- and h- distributions

Distributions	Robust scale estimators	
	MAD_n	T_n
g = 0.0 and h = 0.0	0.0021	0.0040
g = 0.5 and h = 0.0	0.0327	0.0286
g = 1.0 and h = 0.0	0.0255	-0.0035

The Type I error rates for each investigated test was determined using 5000 simulated datasets for 0.05 statistical significance level (α =0.05).This 5000 dataset is chosen due to Manly's (1997) observation. He found 5000 datasets has better sampling limits within which estimated significance levels will fall 99% of the time were obtained when compared to the use of 1000 data sets. Furthermore, any value more than 5000 data sets will not produce large difference on the Type I error rates. To verify this claim, a few additional simulations with 8000 and 10,000 data sets are tested and the results did not differ from 5000 data sets. For example, the Type I error for 5000, 8000 and 10,000 data sets using Parametric Bootstrap test with MAD_n are 0.0366, 0.0366 and 0.0359. The differences between this Type I error rates are significantly low (less than 0.001). After that, each of these simulated datasets will be bootstrapped 599 times with the group means were (0, 0) (refer to Section 3.6).

3.6 Bootstrap Method

Percentile bootstrap method is used in this study to test the hypothesis in proposed procedure. Wilcox & Keselman (2010) stated in their study that percentile bootstrap method generally has a practical advantage when using measures of location that are relatively insensitive to outliers. One type of robust measures of location is trimming that based on the shape of distribution using *MOM* estimator.

Thus, the percentile bootstrap was chosen to obtain the *p*-values of the T_{N0} statistics. Md Yusof *et al.* (2010) referred to some steps to obtain the *p*-value of the T_1 statistics by using the percentile bootstrap method. Hence, this study had used those steps as a guideline to obtain the *p*-value for the procedures under T_{N0} statistics. The steps are as following:

Step 1: Based on the available data, calculate the modified Parametric Bootstrap test statistics (T_{N0}).

Step 2: Generate bootstrap samples by randomly sampling with replacement n_j observations from the j^{th} group yielding $X^*_{(1)j}$, $X^*_{(2)j}$,...,$X^*_{(nj)j}$.

Step 3: Each of the sample points in the bootstrapped groups must be centered at their respective estimated trimmed mean so that the sample trimmed mean is zero, such that $C^*_{ij} = X^*_{ij} - \bar{X}_{tj}$, i= 1,2,…,n_j.

Step 4: Let T^*_{NO} be the value of T_{NO} when applied to the C^*_{ij} values.

Step 5: Repeat step 1 to step 4 B times yielding$T^*_{(N0)_1}$,$T^*_{(N0)_2}$, …,$T^*_{(N0)B}$.

Step 6: Calculate the p-value as (number of $T^*_{NOB} > T_{NO}$) / B.

The calculated p-values represents the estimates rates of Type I error for the procedures investigated under the Parametric Bootstrap statistic.

In order to make the variability of the estimated percentile acceptably low, Efron and Tibshirani (1993) in their study recommended that B should be at least 500 or 1000. Hence, to save the running time, we set B= 599 with the reason that the lowest value that can make α a multiple of $(B + 1)^{-1}$ is 599. To support our decision, trials on various number of bootstrap from B = 599 to 999 found that there was little differences on the values of the Type I error. For example, for Parametric Bootstrap test with robust scale estimator MAD_n, the Type I error for 599, 699, 799, 899 and 999 bootstrap samples were 0.0306, 0.0312, 0.0298, 0.0310 and 0.0310, respectively. The differences between these Type I error rates are small. Thus, using B = 599 is the most suitable to use in this study.

3.7 Analysis on Real Data

Next, the performance of the modified Parametric Bootstrap test with MAD_n and T_n as trimming criteria were demonstrated on real data. Two groups from normal data and two groups from non-normal data were chosen. Then, the *p*-values produced by the proposed procedures were compared with the classical parametric methods and non-parametric methods.

CHAPTER FOUR

RESULTS AND ANALYSIS

4.1 Introduction

This chapter is focused on the proposed test statistics which is Parametric Bootstrap test to test the equality of central tendency of independent groups. This statistic was modified by replacing the usual mean and variances with trimmed mean and Winsorized variances, respectively. Parametric Bootstrap test is used with *MOM* as the location measure and MAD_n or T_n as the scale measure. In addition, MAD_n and T_n also used as the scale estimator in the trimming criteria, suggested by Rousseeuw and Croux (1993). These procedures are to be compared with classical parametric test (i.e. independent sample *t*-test and ANOVA test) and non-parametric test (i.e. Mann Whitney test and Kruskal Wallis test) in terms of Type I error rates for their robustness. In order to highlight the strength and weaknesses of each of the procedures, various conditions were considered in this study such as balanced or unbalanced sample sizes, the shapes of distribution, group variances and nature of pairings. For the number of groups, the procedures were tested under two cases namely two ($J = 2$) and three ($J = 3$) groups. The total sample sizes for $J = 2$ is 40 while for $J = 3$, the total sample sizes is 60. The results produced by each procedure in form of Type I error rates are presented in Tables 4.1 – Tables 4.12. Then, performance of all procedures is tested by original data.

In the first column of all tables are the types of distribution. *g*- and *h*- distribution were considered to represent three different level of skewness. They are $g = 0.0$ and $h = 0.0$ for normal distribution, $g = 0.5$ and $h = 0.0$ for moderately skewed distribution, and $g = 1.0$ and $h = 0.0$ for extremely skewed distribution. The second column for

homogeneous variances cases represents the group variances. On the other hand, for heterogeneous variances cases, the second column represents the nature of pairings of the sample sizes and group variances namely positive pairings (P) and negative pairings (N). The rest of the columns show the Type I error rates produced by the Parametric Bootstrap test with MAD_n and T_n, the classical parametric test and the non-parametric test. The rows represent the "Average" Type I error rate of each procedure corresponding to each distributional shape. The last row of every table displays the "Grand Average" values which represented the overall performance of each procedure by averaging all of Type I error rates produced by each investigated procedure.

Bradley's liberal criterion is adopted to determine the robustness of the test under different conditions especially a test that insensitive to non-normal distribution and heterogeneous variances. Based on this criterion, a test is considered to be robust with respect to Type I error if the empirical rate of Type I error (α) contained in the interval from 0.5α to 1.5α. On the other hand, a test is considered to be non-robust if the empirical Type I error rate straying outside this interval. In this study, the criterion of significant will be set at α=0.05. Thus, a test that produced the empirical rates within the interval of 0.025 to 0.075 can be considered robust. That is, in the conditions of non-normality and unequal variances, if the empirical rate of Type I error is contain in this interval, a test can be considered insensitive to the violation of the assumptions. Empirical Type I error rate below than 0.025 level is considered conservative while those above the 0.075 level is considered liberal.

Apart from that, according to Guo and Luh (2000), the investigated procedure is considered robust if the empirical Type I error rates straying below 0.075 levels.

However, in this study Bradley's liberal criterion is chosen since this robust criterion was widely used by most recent robust statistics researchers in judging robustness (e.g. Cribbie *et al.,* 2012; Md Yusof *et al*., 2012a; Othman *et al*., 2004; Wilcox & Keselman, 2003; Keselman, *et al.*, 2002; Guo & Luh, 2000). A procedure that able to produce the Type I error rates closest the significant level of 0.05 is considered the best procedure.

4.2 New Parametric Bootstrap Procedure

Parametric Bootstrap test was introduced by Krishnamoorthy *et al*. (2007) in their study. Then, Cribbie *et al*.(2012) adopted this test using trimmed mean as the central tendency measures. In this study, Parametric Bootstrap test were modified by using new trimming strategy which use popular robust scale estimators namely MAD_n and T_n as trimming criteria.

The analysis on the Type I error is organized into two cases of groups (J = 2 and J = 3). For each case, the analysis on the Type I error covers the two types of sample sizes (balanced and unbalanced). Then, the Type I error is obtain by combining the Parametric Bootstrap test with percentile bootstrap method.

4.2.1 Type I Error (J = 2)

The results of the analysis on the Type I error rates for J = 2 using Parametric Bootstrap test, independent sample *t*-test and Mann Whitney test are shown in Table 4.1 to Table 4.6. The empirical Type I error rates are displayed for balanced and unbalanced design. For each table, the values which satisfy the Bradley's liberal criterion are highlighted in bold. The "Average" and "Grand Average" values that satisfied the criterion are also highlighted in bold.

4.2.1.1 Balanced sample sizes and homogeneous variances

The first condition used in this study is balanced sample sizes and homogeneous variances. The empirical Type I error rates for this condition is displayed in Table 4.1. Here, the aim is to identify those procedures that are able to control the Type I error rates within the 0.025 to 0.075 interval.

By referring to Bradley's robust criterion, all the Type I error rates under PB test with MAD_n and T_n as trimming criteria across three types of distributions are conservative, ranging from 0.0192 to 0.0240. Therefore, the results for both new trimming strategies are generally not robust with regard to Bradley's liberal criterion.

In contrast to PB test, the results for both independent sample *t*-test and Mann Whitney test fall within the 0.025 to 0.075 interval. The "Average" values in the last row of the table shows that Mann Whitney test (0.0539) produces the best Type I error rates followed by independent sample *t*-test (0.0453).

Next, the Type I error rates are evaluated with respect to distributional shape. For the normal distribution, independent sample *t*-test generates the best result while Mann Whitney test performs better when the distribution is extremely skewed. Based on this table, all the procedures showed an inverse relationship between the Type I error rates and the level of skewness. Therefore, the rates of the Type I error decreased as the level of skewness increased.

Table 4.1

Type I error rates for balanced sample sizes and homogeneous variances

n = (20, 20)					
Distribution Shape	Variances (Equal)	Test statistics			
		PB test with MAD_n	PB test with T_n	*t*-test	Mann Whitney test
g= 0.0 h=0.0	1, 1	0.0214	0.0240	**0.0528**	**0.0566**
g= 0.5 h=0.0	1, 1	0.0212	0.0224	**0.0474**	**0.0526**
g= 1.0 h=0.0	1, 1	0.0192	0.0206	**0.0358**	**0.0526**
Average		0.0206	0.0223	**0.0453**	**0.0539**

4.2.1.2 Balanced sample sizes and heterogeneous variances (moderate)

The empirical Type I error rates were obtained from the tests conducted on groups having equal number of observations and two types of variance heterogeneity namely moderately unequal variances and largely unequal variances. Table 4.2 displays the empirical Type I error rates for balanced sample sizes and moderately unequal variances.

For case of heterogeneous variances, the second column represents the nature of pairing. For case of balanced sample sizes, positive pairing (P) resulted from the association of the balanced observations with the lowest and the highest variances. In contrast to positive pairing, negative pairing (N) resulted when the group having the balanced observations associated with the highest and the lowest variances. Then the Type I error rates for both pairings were averaged and recorded under "Average" for each type of distribution.

Table 4.2

Type I error rates for balanced sample sizes and heterogeneous variances (moderate)

n = (20, 20)					
Distribution Shape	Variances (1,8)	Test statistics			
		PB test with MAD_n	PB test with T_n	*t*-test	Mann Whitney test
g= 0.0 h=0.0	P	**0.0286**	**0.0296**	**0.0586**	0.0840
	N	**0.0254**	**0.0264**	**0.0522**	0.0786
Average		**0.0270**	**0.0280**	**0.0554**	0.0813
g= 0.5 h=0.0	P	**0.0362**	**0.0398**	**0.0670**	**0.0720**
	N	**0.0358**	**0.0326**	**0.0590**	**0.0646**
Average		**0.0360**	**0.0362**	**0.0630**	**0.0683**
g= 1.0 h=0.0	P	**0.0508**	**0.0496**	0.1030	0.0792
	N	**0.0472**	**0.0444**	0.0932	**0.0726**
Average		**0.0490**	**0.0470**	0.0981	0.0759
Grand Average		**0.0373**	**0.0371**	**0.0722**	0.0752

All the Type I error rates for PB test with MAD_n and T_n satisfied the Bradley's liberal criterion of robustness (0.025 to 0.075). For independent sample *t*-test, all the values fall within the interval except for the extremely skewed distribution where the Type I error rates becomes liberal with rates ranging from 0.0932 to 0.1030. For Mann Whitney test all their values straying above 0.075 except for moderately skewed distribution and for extremely skewed distribution with negative pairing (N).

The overall result of all the procedures across the three types of distribution which represented by "Grand Average" shows that PB test with MAD_n and T_n and independent sample *t*-test were found to be robust. In fact, the average Type I error

rates for the PB test with MAD_n (0.0373) is the closest to the significant level (0.05) followed by the T_n procedure (0.0371).

All these procedures are robust in distributional shape when tested under normal distribution except for Mann Whitney test. For the moderately skewed distribution, all procedures are robust with their average Type I error ranging from 0.0360 to 0.0683. The average Type I error for the extremely skewed distribution showed that PB test with new trimming criteria are considered robust in accordance with Bradley's robustness criterion. The results from MAD_n and T_n procedures indicated that it functions effectively under extremely skewed distribution with moderate unequal variances.

4.2.1.3 Balanced sample sizes and heterogeneous variances (large)

The empirical Type I error rates obtained from the test performed on the condition of balanced sample sizes and largely unequal variances are showed in Table 4.3.

The results from the table indicate that all the empirical Type I error rates for PB test with new trimming criteria namely MAD_n and T_n were acceptable with respect to Bradley's robustness criterion across the three types of distributions except for the negative pairing (N) under normal distribution. However, the "Average" Type I error rates under normal distribution for these procedures fall within the interval. For independent sample *t*-test, the Type I error rates under extremely skewed distribution becomes very liberal exceeded 0.1 levels. Apart from that, none of the Type I error rates from Mann Whitney test falls within the 0.025 to 0.075 interval. The values ranging from 0.0794 to 0.0938 meaning that this test should be considered not robust.

Table 4.3

Type I error rates for balanced sample sizes and heterogeneous variances (large)

n = (20, 20)					
Distribution Shape	Variances (1,36)	Test statistics			
		PB test with MAD_n	PB test with T_n	*t*-test	Mann Whitney test
g= 0.0 h=0.0	P	**0.0294**	**0.0294**	**0.0618**	0.0938
	N	0.0232	0.0224	**0.0554**	0.0926
Average		**0.0263**	**0.0259**	**0.0586**	0.0932
g= 0.5 h=0.0	P	**0.0412**	**0.0436**	0.0788	0.0862
	N	**0.0358**	**0.0344**	**0.0700**	0.0794
Average		**0.0385**	**0.0390**	**0.0744**	0.0828
g= 1.0 h=0.0	P	**0.0568**	**0.0572**	0.1292	0.0854
	N	**0.0572**	**0.0524**	0.1194	0.0804
Average		**0.0570**	**0.0548**	0.1243	0.0829
Grand Average		**0.0406**	**0.0399**	0.0858	0.0863

The "Grand Average" which represents the overall performance of all the procedure across the three types of distributions displayed on the last row of the table. Among all of these procedures, PB test with MAD_n and T_n were found to be robust. Yet again, PB test with MAD_n (0.0406) establish itself as the best procedure with its Type I error rate is closest to the significant level followed by T_n procedure (0.0399).

Across distributional shapes, independent sample *t*-test performs better when the distribution is normal while for the moderately skewed distribution, PB test with T_n generate the best result. For the extremely skewed distribution, MAD_n (0.0570) and

T_n (0.0548) procedures show robust average value with T_n procedure generates the best result.

4.2.1.4 Unbalanced sample sizes and homogeneous variances

Next, the empirical Type I error rates produced by each procedure for the condition of unbalanced sample sizes and homogeneous variances was examined and shown in Table 4.4.

Table 4.4

Type I error rates for unbalanced sample sizes and homogeneous variances

n = (15, 25)					
Distribution Shape	Variances (Equal)	Test statistics			
		PB test with MAD_n	PB test with T_n	*t*-test	Mann Whitney
g= 0.0 h=0.0	1, 1	0.0244	0.0236	**0.0490**	**0.0510**
g= 0.5 h=0.0	1, 1	0.0234	**0.0262**	**0.0468**	**0.0502**
g= 1.0 h=0.0	1, 1	0.0196	0.0208	**0.0382**	**0.0502**
Average		0.0225	0.0235	**0.0447**	**0.0505**

The 'Average' values on the last row of the table represent the overall performance of the investigated procedures. Based on these values, PB test with MAD_n and T_n as trimming criteria produced conservative Type I error rates, ranging from 0.0225 to 0.0235. In fact, it was observed that none of the Type I error rates recorded by PB test with MAD_n fall within the interval. All their values become conservative, straying below the level of 0.025. Therefore, PB test with MAD_n was not robust under this condition. For T_n procedure, all the Type I error rates fall outside the Bradley's interval except for the value under moderately skewed distribution. On the

other hand, independent sample *t*-test and Mann Whitney test provided an excellent Type I error control where the "Average" values produced by these procedures fulfill the Bradley's liberal criterion. In fact, the best procedure for this condition is Mann Whitney test (0.0505) with its Type I error rate is closest to the significant level followed by the independent sample *t*-test (0.0447).

Across distributional shapes, independent sample *t*-test and Mann Whitney test is robust under symmetric distribution and Mann Whitney test perform better under the skewed distributions. There is also an inverse relationship between the Type I error rates and the level of skewness except for the PB test with T_n procedure. The Type I error rates for T_n procedure suddenly increase from 0.0236 to 0.0262 when the distribution is moderately skewed.

4.2.1.5 Unbalanced sample sizes and heterogeneous variances (moderate)

Table 4.5 displays the empirical Type I error rates obtained from the test perform on the unbalanced sample sizes and moderately unequal variances.

For case of heterogeneous variances with unbalanced sample sizes, the second column represents the nature of pairing where the positive pairing (P) resulted from the association of the lowest variances with the smallest observation and the association of the highest variance with the largest observation. In contrast to positive pairing, negative pairing (N) resulted when the group having the smallest observation associated with the highest variances and the largest observation associated with the lowest variances. Then the Type I error rates for both pairings were averaged and recorded under "Average" for each type of distribution.

Table 4.5

Type I error rates for unbalanced sample sizes and heterogeneous variances (moderate)

n = (15, 25)					
Distribution Shape	Variances (1,8)	Test statistics			
		PB test with MAD_n	PB test with T_n	*t*-test	Mann Whitney test
g= 0.0 h=0.0	P	**0.0278**	**0.0308**	0.0246	**0.0500**
	N	**0.0274**	**0.0264**	0.1052	0.1064
Average		**0.0276**	**0.0286**	**0.0649**	0.0782
g= 0.5 h=0.0	P	**0.0316**	**0.0344**	**0.0346**	**0.0426**
	N	**0.0382**	**0.0396**	0.1222	0.0874
Average		**0.0349**	**0.0367**	0.0784	**0.0650**
g= 1.0 h=0.0	P	**0.0366**	**0.0422**	**0.0606**	**0.0472**
	N	**0.0612**	**0.0606**	0.1498	0.0980
Average		**0.0489**	**0.0514**	0.1052	**0.0726**
Grand Average		**0.0371**	**0.0389**	0.0828	**0.0719**

The overall result of all the procedures across all the three different level of skewness which represented by the "Grand Average" on the last row of the table was examined. It was observed that with the exception of the independent sample *t*-test, all the other procedures were found to be robust as their values fulfill the Bradley's liberal criterion of robustness. Moreover, all the results for PB test with new trimming strategies namely MAD_n and T_n never strayed outside the Bradley's robust criterion. Based on the 'Grand Average' values, PB test with T_n (0.0389) is the best procedure where the Type I error rates for this test is the closest to the significant level, followed by PB test with MAD_n (0.0371) and Mann Whitney test (0.0719).

For the case of independent sample *t*- test, the "Average" value under normal distribution falls within the interval. However, the "Average" value for this test becomes liberal with the rates straying above 0.075 when the distribution is skewed. In contrast to the independent sample *t*-test, although Mann Whitney test showed a good control of Type I error rates under skewed distribution, but this test failed to perform well under normal distribution with its Type I error rates fall outside the interval.

With regard to distributional shape, all the procedures show robust "Average" values ranging from 0.0276 to 0.0646 except the Mann Whitney test for symmetric distribution. For the skewed distributions, all the procedures show robust "Average" values ranging from 0.0349 to 0.0719 except the independent sample *t*-test.

From this table, empirically the positive and negative pairings for the independent sample *t*-test and Mann Whitney test typically produce conservative and liberal results, respectively. The negative pairings, which refer to the association of the smallest observations with the largest variances and the association of the largest observations with the smallest variances, generates higher Type I error rates exceeding 0.075 levels compared to the positive pairings.

4.2.1.6 Unbalanced sample sizes and heterogeneous variances (large)

Empirical Type I error rates for unbalanced sample sizes and largely unequal variances is presented in Table 4.6.

Table 4.6

Type I error rates for unbalanced sample sizes and heterogeneous variances (large)

n = (15, 25)					
Distribution Shape	Variances (1,36)	Test statistics			
		PB test with MAD_n	PB test with T_n	*t*-test	Mann Whitney test
g= 0.0 *h*=0.0	P	**0.0284**	**0.0312**	0.0198	**0.0508**
	N	0.0232	**0.0252**	0.1268	0.1244
Average		**0.0258**	**0.0282**	**0.0733**	0.0876
g= 0.5 *h*=0.0	P	**0.0370**	**0.0418**	**0.0324**	**0.0440**
	N	**0.0358**	**0.0402**	0.1486	0.1068
Average		**0.0364**	**0.0410**	0.0905	0.0754
g= 1.0 *h*=0.0	P	**0.0500**	**0.0566**	**0.0722**	**0.0446**
	N	**0.0632**	**0.0646**	0.2054	0.1072
Average		**0.0566**	**0.0606**	0.1388	0.0759
Grand Average		**0.0396**	**0.0433**	0.1009	0.0796

As shown in the table, all the "Average" values for PB test with new trimming criteria namely MAD_n and T_n across the three types of distribution fulfill the Bradley's liberal criterion. For normal distribution, the "Average" value for independent sample *t*-test falls within the interval. The "Average" value for independent sample *t*-test worsens when the distribution becomes skewed. Apart from that, none of the "Average" values for Mann Whitney test acceptable with regard to Bradley's robust criterion.

Based on the "Grand Average" values that represent the overall performance of all procedures, PB test competes well with MAD_n and T_n because the Type I error rates

recorded for these procedures are within 0.025 to 0.075. Therefore, both of these procedures are robust under Bradley's robust criterion. PB test with T_n generated "Grand Average" closest to the 0.05 significant level with a value of 0.0433 followed by PB test with MAD_n with a value of 0.0396.

All the procedures except Mann Whitney test showed robust average Type I error rates with T_n procedure (0.0282) emerging as the best procedure for normal distribution in distributional shape. The average values for the skewed distribution showed that PB test with MAD_n and T_n as trimming criteria are considered robust in accordance with Bradley's liberal criterion. These new trimming strategies perform better than independent sample *t*-test and Mann Whitney test under extremely skewed distribution with unequal variances and unbalanced sample sizes. From this table, it could be comprehended that empirically the positive and negative pairings for the independent sample *t*-test and Mann Whitney test typically produce conservative and liberal results respectively, which the negative pairings generate higher Type I error rates exceeded 0.075 level compared to the positive pairings.

4.2.2 Type I Error (J = 3)

Apart from the case of J = 2, this study also covered the analysis of the Type I error rates for the case of J = 3 using Parametric Bootstrap test, ANOVA test and Kruskal Wallis test. The previous section already discussed on the results of J = 2 case in terms of Type I error rates. The condition for the case of J = 3 are the same except for changes in the total of sample sizes to N = 60. Throughout this section, the results for the J = 3 case are presented in Table 4.7 to Table 4.12. As mentioned earlier, the values that fulfilled the Bradley's liberal criterion of robustness were highlighted in bold.

4.2.2.1 Balanced sample sizes and homogeneous variances

Table 4.7 shows the empirical Type I error rates produced by each procedure for the condition of balanced sample sizes and homogeneous variances under three groups case.

Table 4.7

Type I error rates for balanced sample sizes and homogeneous variances

n = (20, 20, 20)					
Distribution Shape	Variances (Equal)	Test statistics			
		PB test with MAD_n	PB test with T_n	ANOVA test	Kruskal wallis test
g= 0.0 h=0.0	1,1,1	0.0154	0.0172	**0.0532**	**0.0486**
g= 0.5 h=0.0	1,1,1	0.0140	0.0154	**0.0510**	**0.0532**
g= 1.0 h=0.0	1,1,1	0.0110	0.0148	**0.0412**	**0.0532**
Average		0.0135	0.0158	**0.0485**	**0.0517**

First, the new trimming strategies using PB test produce conservative Type I error rates. Thus, these procedures can be considered not robust for this condition. On the other hand, all the Type I error rates for ANOVA test and Kruskal Wallis test are satisfy the Bradley's robust criterion with their values fall within the 0.025 to 0.075 interval. Therefore, these procedures were found to be robust.

The last row of the table showed the "Average" values that represent the overall performances of all the procedures. It was observed that the Type I error rates produced by ANOVA test (0.0485) is closest to the 0.05 level followed by Kruskal Wallis test (0.0517) while PB test with MAD_n (0.0135) recorded the lowest value.

Thereafter, the Type I error rates are evaluated with respect to distributional shape. Kruskal Wallis test generate the best result for symmetric distribution and extremely skewed distribution while ANOVA test performs better when the distribution is moderately skewed. There is also an inverse relationship between the Type I error rates and the level of skewness except for Kruskal Wallis procedure. The Type I error rates for this procedure suddenly increases from 0.0486 to 0.0532 when the distribution is skewed. In general, from this analysis, ANOVA test and Kruskal Wallis test perform better than the new trimming criteria under homogeneous variances with balanced sample sizes.

4.2.2.2 Balanced sample sizes and heterogeneous variances (moderate)

Next, the empirical Type I error rates of the test conducted under the condition of balanced sample sizes and unequal variances are examined. As in the previous section, the Type I error rates is observed under two types of variance heterogeneity that is moderately unequal variances and largely unequal variances. Table 4.8 displays the results for all procedures for balanced sample sizes and moderately unequal variances.

The "Grand Average" which represents the overall performance of the procedures across the distributions shows that all the procedures were found to be robust where the corresponding Type I error rates fulfill the Bradley's robust criterion. Based on these values, PB test with MAD_n (0.0290) is the best procedure since its Type I error rate is nearest to the significant level of 0.05 followed by PB test with T_n (0.0288), Kruskal Wallis test (0.0740) and lastly, ANOVA test (0.0742).

Table 4.8

Type I error rates for balanced sample sizes and heterogeneous variances (moderate)

n = (20, 20, 20)					
Distribution Shape	Variances (1,8,16)	Test statistics			
		PB test with MAD_n	PB test with T_n	ANOVA test	Kruskal wallis test
g= 0.0 h=0.0	P	0.0168	0.0162	**0.0664**	**0.0734**
	N	0.0210	0.0212	**0.0684**	**0.0748**
Average		0.0189	0.0187	**0.0674**	**0.0741**
g= 0.5 h=0.0	P	**0.0252**	**0.0258**	**0.0678**	**0.0678**
	N	**0.0294**	**0.0306**	**0.0698**	**0.0708**
Average		**0.0273**	**0.0282**	**0.0688**	**0.0693**
g= 1.0 h=0.0	P	**0.0414**	**0.0392**	0.0864	0.0764
	N	**0.0402**	**0.0396**	0.0862	0.0810
Average		**0.0408**	**0.0394**	0.0863	0.0787
Grand Average		**0.0290**	**0.0288**	**0.0742**	**0.0740**

Under normal distribution, the new trimming strategies namely Parametric Bootstrap test with MAD_n and T_n produce the conservative Type I error rates ranging from 0.0162 to 0.0212 while ANOVA test and Kruskal Wallis test successfully control the Type I error rates within the 0.025 to 0.075 interval. However, as the level of skewness increases to the extreme level, the Type I error rates for MAD_n and T_n procedures increases and fall within the interval. In contrast to PB test, ANOVA test and Kruskal Wallis test failed to control the Type I error rates under extremely skewed distribution with its value straying above 0.075. In general, the new trimming strategies perform better than ANOVA test and Kruskal Wallis test under extremely

skewed distribution with unequal variances and MAD_n procedure (0.0408) generate the best result for this type of distribution.

4.2.2.3 Balanced sample sizes and heterogeneous variances (large)

The results of all procedures in terms of Type I error rates for balanced sample sizes and largely unequal variances are shown in Table 4.9.

Table 4.9

Type I error rates for balanced sample sizes and heterogeneous variances (large)

n = (20, 20, 20)					
Distribution Shape	Variances (1,1,36)	Test statistics			
		PB test with MAD_n	PB test with T_n	ANOVA test	Kruskalwallis test
g= 0.0 h=0.0	P	0.0166	0.0174	0.0858	0.0838
	N	0.0154	0.0150	0.0874	0.0892
Average		0.0160	0.0162	0.0866	0.0865
g= 0.5 h=0.0	P	0.0230	0.0238	0.1054	0.0786
	N	0.0230	0.0224	0.1026	0.0818
Average		0.0230	0.0231	0.1040	0.0802
g= 1.0 h=0.0	P	**0.0306**	**0.0308**	0.1600	0.0826
	N	**0.0358**	**0.0336**	0.1502	0.0894
Average		**0.0332**	**0.0322**	0.1551	0.0860
Grand Average		0.0241	0.0233	0.1152	0.0842

By referring to the "Grand Average" values that represent the overall performance of all procedures, none of the Type I error rates falls within the 0.025 to 0.075 interval. Therefore, none of the procedures can be considered robust under Bradley's liberal

criterion. However, the "Grand Average" values for PB test with MAD_n (0.0241) and T_n (0.0233) are slightly conservative with regard to Bradley's robustness criterion. Among the non-robust procedures, the PB test with MAD_n produce the best Type I error rates followed by T_n procedure.

PB test with MAD_n and T_n provide a good control of the Type I error rates under extremely skewed distribution with their values fall within the 0.025 to 0.075 interval. In contrast, all the Type I error rates for ANOVA test and Kruskal Wallis test became liberal exceeding 0.075 levels.

With respect to distributional shape, none of the Type I error rates for all procedures fall within the interval under symmetric distribution. This situation is same for moderately skewed distribution where all the average values produced by the investigated procedures straying outside the bound of robustness. For extremely skewed distribution, only PB test with MAD_n and T_n showed some improvement where its Type I error rates slightly increase and satisfy the Bradley's robust criterion. Therefore, the new trimming strategies using PB test showed better Type I error control under extremely skewed distribution with unequal variances compared to ANOVA test and Kruskal Wallis test. In fact, the PB test with MAD_n produced 'Grand Average' value slightly below the lowest limit of the Bradley's interval with the difference of 0.0009.

4.2.2.4 Unbalanced sample sizes and homogeneous variances

The performance of the investigated procedures for the condition of unbalanced sample sizes and homogeneous variances are displayed in Table 4.10.

Based on the 'Grand Average" values displayed on the last row of the table which represents the overall performance of all the procedures shows that ANOVA test and Kruskal Wallis test provided an excellent control of Type I error rates where their values are consistent and close to the significant level of 0.05. In contrast, PB test with MAD_n and T_n failed to perform well under this condition with none of their Type I error rates fulfill the Bradley's liberal criterion.

Table 4.10

Type I error rates for unbalanced sample sizes and homogeneous variances

n = (15, 20, 25)					
Distribution Shape	Variances (Equal)	Test statistics			
		PB test with MAD_n	PB test with T_n	ANOVA test	Kruskal wallis test
g= 0.0 h=0.0	1,1,1	0.0164	0.0176	**0.0510**	**0.0500**
g= 0.5 h=0.0	1,1,1	0.0144	0.0174	**0.0486**	**0.0510**
g= 1.0 h=0.0	1,1,1	0.0114	0.0138	**0.0402**	**0.0510**
Average		0.0141	0.0163	**0.0466**	**0.0507**

Across distributional shapes, Kruskal Wallis test generate the best result for symmetric distribution and skewed distributions. There is also an inverse relationship between the Type I error rates and the level of skewness except for Kruskal Wallis procedure. The Type I error rates for this procedure suddenly increases from 0.0500 to 0.0510 when the distribution is skewed. In general, from this analysis, ANOVA test and Kruskal Wallis test perform better than the new trimming criteria under homogeneous variances and unbalanced sample sizes.

4.2.2.5 Unbalanced sample sizes and heterogeneous variances (moderate)

Table 4.11 displays the empirical Type I error rates for unbalanced sample sizes and moderately unequal variances. For this case, this study also encompassed the investigation on the positive pairings (P) and negative pairing (N). For each distribution, the Type I error rates for both pairing were averaged and recorded as "Average".

Table 4.11

Type I error rates for unbalanced sample sizes and heterogeneous variances (moderate)

n = (15, 20, 25)					
Distribution Shape	Variances (1, 8, 16)	Test statistics			
		PB test with MAD_n	PB test with T_n	Anova test	Kruskal wallis test
g= 0.0 h=0.0	P	0.0164	0.0180	**0.0378**	**0.0468**
	N	0.0184	0.0182	0.1044	0.1008
Average		0.0174	0.0181	**0.0711**	**0.0738**
g= 0.5 h=0.0	P	0.0226	**0.0262**	**0.0402**	**0.0466**
	N	**0.0274**	**0.0290**	0.1082	0.0904
Average		**0.0250**	**0.0276**	**0.0742**	**0.0685**
g= 1.0 h=0.0	P	**0.0330**	**0.0338**	**0.0546**	**0.0560**
	N	**0.0412**	**0.0406**	0.1194	0.1012
Average		**0.0371**	**0.0372**	0.0870	0.0786
Grand Average		**0.0265**	**0.0276**	0.0774	**0.0736**

The "Grand Average" values across three types of distributions which represent the overall results displayed on the last row of the table. Based on these values, it was shown that with the exception of the ANOVA test, all the other procedures are robust

with their Type I error rates straying within the 0.025 to 0.075 interval. PB test with T_n (0.0276) can be considered as the best procedure with its Type I error rate being closest to the significant level followed by PB test with MAD_n (0.0265) and Kruskal Wallis test (0.0736).

With regard to distributional shape, ANOVA test and Kruskal Wallis test showed a good control of Type I error rates under normal distribution and moderately skewed distribution. The "Average" values produced by these procedures fall within the interval. Under extremely skewed distribution, the modified procedures using new trimming criteria seems to have a better control of Type I error rates compared to ANOVA test and Kruskal Wallis test with the "Average" values for MAD_n and T_n fall within the interval even though all their Type I error rates under symmetric distribution becomes conservative. Apart from that, the negative pairings for ANOVA test and Kruskal Wallis test generates higher Type I error rates (above 0.075) compared to that of positive pairings.

4.2.2.6 Unbalanced sample sizes and heterogeneous variances (large)

Lastly, the performance of all investigated procedures was examined under the condition of unbalanced sample sizes and largely unequal variances. Table 4.12 displays the empirical Type I error rates produced by new trimming strategies using PB test, ANOVA test and Kruskal Wallis test.

It was observed that only PB test with T_n (0.0257) is robust with its "Grand Average" value on the last row of the table satisfied the Bradley's robust criterion. All the other procedures were not robust with their Type I error rates fall outside the interval. The

"Grand Average" value represents the overall performance of the investigated procedures across the three types of distributions.

Table 4.12

Type I error rates for unbalanced sample sizes and heterogeneous variances (large)

n = (15, 20, 25)					
Distribution Shape	Variances (1,1,36)	Test statistics			
		PB test with MAD_n	PB test with T_n	ANOVA test	Kruskal wallis test
g= 0.0 h=0.0	P	0.0148	0.0146	**0.0414**	**0.0614**
	N	0.0158	0.0178	0.1514	0.1106
Average		0.0153	0.0162	0.0964	0.0860
g= 0.5 h=0.0	P	0.0194	0.0226	**0.0620**	**0.0568**
	N	**0.0262**	**0.0280**	0.1696	0.1012
Average		0.0228	**0.0253**	0.1158	0.0790
g= 1.0 h=0.0	P	0.0226	**0.0276**	0.1094	**0.0614**
	N	**0.0416**	**0.0436**	0.2172	0.1048
Average		**0.0321**	**0.0356**	0.1633	0.0831
Grand Average		0.0234	**0.0257**	0.1252	0.0827

With respect to distributional shape, the average Type I error rates for PB test with MAD_n and T_n as trimming criteria becomes highly conservative, ranging from 0.0146 to 0.0178 under the symmetric distribution. For both ANOVA test and Kruskal Wallis test, none of their "Average" values fulfill the Bradley robust criterion for all three types of distributions. Some of their Type I error rates become very liberal exceeded 0.1 levels. Apart from that, only PB test with T_n showed a good control of Type I error rates under moderately skewed distribution with its "Average" value is

0.0253. For extremely skewed distribution, PB test with MAD_n and T_n produced a better Type I error control with their "Average" values fall within the Bradley's interval (0.025-0.075). Thus, these new trimming criteria were proven to provide a better result compared to ANOVA test and Kruskal Wallis test for extremely skewed distribution with unequal variances and unbalanced sample sizes.

With regard to the design types, the Type I error rates for the $J = 2$ case seems higher and better than for the case of $J = 3$ indicating that $J = 2$ procedures are better in controlling the Type I error rates.

4.3 Analysis on Real Data

The performance of the Parametric Bootstrap test with MAD_n and T_n as trimming criteria were demonstrated on real data. These procedures were compared with the classical parametric method and non-parametric method. Two classes (groups) of standard 6 for subject 'Pendidikan Kesihatan' and 'Pendidikan Seni' were chosen. For subject 'Pendidikan Kesihatan', the sample sizes for Class 1 and Class 2 were 29 and 16, respectively. While for subject 'Pendidikan Seni', the sample sizes for Class 1 and Class 2 were 29 and 19, respectively. The marks and descriptive statistic for each subject are given in Table 4.13, Table 4.14 and Table 4.15.

Table 4.13

Marks for each subject

Subject	Class	Data
'Pendidikan Kesihatan'	Class 1	64 66 62 68 60 58 58 56 56 60 48 60 70 72 60 74 72 58 72 64 56 44 50 50 54 54 50 52 44
	Class 2	80 60 60 60 65 60 60 50 52 55 45 55 50 45 42 40
'Pendidikan Seni'	Class 1	65 40 65 40 40 60 40 40 65 75 85 40 45 50 75 75 85 80 50 50 65 85 85 85 50 55 60 80 45
	Class 2	65 70 80 60 65 70 65 55 65 55 60 60 60 45 70 50 50 60 50

Table 4.14

Descriptive Statistic for 'Pendidikan Kesihatan'

Class	Sample Size (*N*)	Mean	Variance	Std. Deviation	Minimum	Maximum
1	29	59.03	71.034	8.428	44	74
2	16	54.94	100.196	10.010	40	80
Total	45	113.97	171.23	18.438	40	80

Table 4.15

Descriptive Statistic for 'Pendidikan Seni'

Class	Sample Size (*N*)	Mean	Variance	Std. Deviation	Minimum	Maximum
1	29	61.21	283.313	16.832	40	85
2	19	60.79	75.731	8.702	45	80
Total	48	61	179.522	12.767	40	85

The Shapiro-Wilk test is employed with a view to determine the normality of the data. Shapiro and Wilk (1965) in their study stated that this test is comparatively quite sensitive to a wide range of non-normality and suitable for small sample sizes even for (n<20). Table 4.16 presents the results of normality for each Class. Based on this table, for subject 'Pendidikan Kesihatan' the *p*-values for Class 1 and Class 2 are greater than critical value (0.05) meaning that the data for both Classes comes from a normal distribution. On the other hand, for subject 'Pendidikan Seni', the *p*-value for Class 1 is 0.005, less than 0.05. Thus, the null hypothesis is rejected and it could be concluded that the data for Class 1 comes from a non-normal distribution.

Table 4.16

Shapiro-Wilk test for normality assumption

		Shapiro –Wilk		
Subject	Class	Statistic	Df	Sig.
'Pendidikan Kesihatan'	1	0.965	29	0.442
	2	0.931	16	0.255
'Pendidikan Seni'	1	0.886	29	0.005
	2	0.962	19	0.609

Then, the Levene's test (1960) is conducted to test the assumption of homoscedasticity for Class 1 and Class 2. Based on Table 4.17, the *p*-value for subject 'Pendidikan Kesihatan' is greater than 0.05. This result is taken as evidence that the assumption has not been violated (i.e. the data for 'Pendidikan Kesihatan' comes from equal variance). In contrast, the data for 'Pendidikan Seni' comes from unequal variance since its *p*-value is less than 0.05.

Table 4.17

Levene's test for homoscedasticity assumption

Levene's Test		
Subject	F	Sig.
'Pendidikan Kesihatan'	0.251	0.619
'Pendidikan Seni'	15.488	0.000

Based on the results from Shapiro-Wilk test and Levene's test, the data for Class 1 and Class 2 (Pendidikan Kesihatan) were found to be normally distributed and equal variances while the data for Class 1 and Class 2 (Pendidikan Seni) were found to be non-normally distributed with unequal variances. To evaluate the performance of the Parametric Bootstrap test with MAD_n and T_n as trimming criteria, the *p*-values of the procedure was recorded. The results for the proposed procedure, independent sample *t*-test and Mann Whitney test for subject 'Pendidikan Kesihatan' and 'Pendidikan Seni' are showed in Table 4.18 and Table 4.19.

Table 4.18

p-values for 'Pendidikan Kesihatan' (normal data and equal variances)

Methods	*p*-value
PB test with MAD_n	0.1276
PB test with T_n	0.0944
t-test	0.1516
Mann Whitney test	0.1632

For 'Pendidikan Kesihatan'subject, all procedures failed to reject the null hypothesis indicated that there was no significant difference in terms of marks between Class 1 and Class 1. However, Parametric Bootstrap test with T_n showed the lowest *p*-value that less than 0.1 level which is 0.0944. Others procedures produced results where *p*-value exceeded 0.1 level. This finding denoted that the proposed procedure performed better than independent sample *t*-test and Mann Whitney test for normal data and equal variances.

Table 4.19

p-values for 'Pendidikan Seni' (non-normal and unequal variances)

Methods	*p*-value
PB test with MAD_n	0.6951
PB test with T_n	0.7172
t-test	0.9110
Mann Whitney test	0.9240

Based on the Table 4.19, the *p*-values for all the procedures used in this study are greater than significant level (0.05) indicated that all the procedures failed to reject null hypothesis. It means that there was no significant difference between Class 1 and Class 2 in terms of their marks. However, the Parametric Bootstrap test with

MAD_n and T_n showed better detection compared to independent sample *t*-test and Mann Whitney test with their *p*-values are 0.6951 and 0.7172. This result indicated that the proposed test is also suitable for non-normal data and unequal variances.

CHAPTER FIVE
CONCLUSIONS

5.1 Introduction

Non-normality and variance heterogeneity are two major problems that researchers most commonly encounter in testing the equality of central tendency. Classical parametric test such as ANOVA test and independent sample *t*-test are known to be sensitive to these assumptions. Violation in these assumptions can distort the Type I error rates and consequently the classical parametric test that is used typically provides invalid results. A common recommendation due to non-normality and variance heterogeneity is to use non-parametric procedures and simple transformation. However, non-parametric procedures are known to be less powerful and simple transformation failed to deal directly with outliers. Hence, this study proposed a robust test statistics which is insensitive to these assumptions. By replacing the usual mean and variances using robust measures of location and scale such as trimmed mean and Winsorized variances, respectively, the proposed test offers the best Type I error control under non-normal distribution and unequal variances.

This study focuses on Parametric Bootstrap test for testing the central tendency measures. Parametric Bootstrap test is originally introduced by Krishnamoorthy *et al.* (2007). However, their study only investigated the robustness of each procedure under variances heterogeneity. Therefore, for this study, a modification of a test statistic is proposed, termed as Parametric Bootstrap test, which used trimmed mean obtained by using *MOM* estimator to make this test robust in dealing with non-normality and variances heterogeneity. *MOM* is flexible in handling outliers in a data

with empirically determines the amount of trimming percentages regarding the shape of distributions. For this new trimming strategy, two robust scale estimators namely MAD_n and T_n were used as trimming criteria. These scale estimators were chosen due to their highest possible value of breakdown point (0.5) and bounded influence function (Rousseeuw and Croux, 1993). These two procedures were compared to the most frequently used classical parametric test (i.e. independent sample *t*-test and ANOVA test) and non-parametric test (i.e. Mann Whitney test and Kruskal Wallis test) in terms of the ability to control Type I error rates under non-normal data and unequal variances.

Under the effect of non-normality and variances heterogeneity, the Type I error rates were calculated for each investigated procedure to determine its robustness. The strength and weaknesses of each procedure in testing the equality of central tendency were based on several manipulated variables. For testing the normality effect, three types of distributional shapes representing different level of skewness and kurtosis were used by using *g*- and *h*- distribution. The distribution of $g = 0.0$ and $h = 0.0$ represents normal distribution, $g = 0.5$ and $h = 0.0$ distribution represents moderately skewed distribution while the extremely skewed distribution is represented by the distribution of $g = 1.0$ and $h = 0.0$. Apart from that, moderately unequal variances with 1:8 (1:8:16) ratio and largely unequal variances with 1:36 ratio (1:1:36) were assigned to this study for the heteroscedasticity effect. Also included in this study is the number of groups ($J = 2$ and $J = 3$) with balanced and unbalanced sample sizes. The other variable such as the nature of pairings (positive and negative) was also considered in this study. All the procedures were simulated 5000 times for significant level of 0.05. This study used Percentile Bootstrap method for the Parametric

Bootstrap statistic in order to test the central tendency measures. As for bootstrap method, 599 bootstrap samples were generated for each simulation.

In this study, Bradley's (1978) liberal criterion is adopted to determine the robustness of all the investigated procedures. Based on this criterion, a procedure that produced its empirical Type I error rates within the 0.025 to 0.075 interval for significant level, $\alpha = 0.05$ is considered robust. Then, the Type I error rates for each procedure were examined and compared with the chosen classical parametric test and non-parametric test in order to determine the best procedure. The best procedure will produce the empirical Type I error rates closest to the significant value of 0.05.

5.2 The new Parametric Bootstrap procedures

Two different procedures were proposed and tested for the Type I error rates under Parametric Bootstrap statistic. Then, these procedures were compared with classical parametric test and non-parametric test that were chosen. First, we concluded the results of the empirical Type I error rates under homogeneous and heterogeneous variances as shown in Table 5.1 to Table 5.4. The values for homogeneous variances were computed by taking the mean of the 'Average' values under the condition of equal variances for balanced and unbalanced sample sizes. While for the heterogeneous variances, the values were computed by taking the mean of the 'Grand Average' values under the condition of unequal variances for balanced and unbalanced sample sizes.

According to these tables, the classical parametric method represented by independent sample *t*-test for $J = 2$ and ANOVA test for $J = 3$, while non-parametric

method is represented by Mann Whitney test for $J = 2$ and Kruskal Wallis test for $J = 3$.

Table 5.1

Average empirical Type I error rates for homogeneous variances(J = 2)

Test statistics	Average Type I error rates
PB test with MAD_n	0.0216
PB test with T_n	0.0229
t-test	**0.0450**
Mann Whitney test	**0.0522**

Table 5.2

Average empirical Type I error rates for homogeneous variances(J = 3)

Test statistics	Average Type I error rates
PB test with MAD_n	0.0138
PB test with T_n	0.0161
ANOVA test	**0.0476**
Kruskal Wallis test	**0.0512**

From the average empirical Type I error rates shown in Table 5.1 and Table 5.2, Parametric Bootstrap test with MAD_n and T_n as trimming criteria were found to be not robust under homogeneous variances. In contrast to Parametric Bootstrap test, the classical parametric method and non-parametric method showed a good control of the Type I error for $J = 2$ and $J = 3$ under homogeneous variances across the three types of distribution with their empirical Type I error rates satisfied the Bradley's robustness criterion.

Table 5.3

Average empirical Type I error rates for heterogeneous variances (J = 2)

Test statistics	Average Type I error rates
PB test with MAD_n	**0.0387**
PB test with T_n	**0.0398**
t-test	0.0854
Mann Whitney test	0.0783

Table 5.4

Average empirical Type I error rates for heterogeneous variances (J = 3)

Test statistics	Average Type I error rates
PB test with MAD_n	**0.0258**
PB test with T_n	**0.0307**
ANOVA test	0.0980
Kruskal Wallis test	0.0786

In contrast to homogeneous variances, the Parametric Bootstrap test with new trimming strategy, MAD_n and T_n provides the best Type I error control for $J = 2$ and $J = 3$ under heterogeneous variances as shown in Table 5.3 and Table 5.4. These two procedures were considered robust with their empirical Type I error rates fulfilled the Bradley's robustness criterion when the assumption of homogeneity is violated. The classical parametric method and non-parametric method for $J = 2$ and $J = 3$ are not robust under heterogeneous variances with their empirical Type I error rates contained outside the interval of 0.025 to 0.075.

Therefore, it can be comprehended that the Parametric Bootstrap test with MAD_n and T_n as robust scale estimators showed a good Type I error control under the effect of

variances heterogeneity even though it produced conservative Type I error rates under homogeneous variances.

Finally, the effects of distributional shapes under heterogeneous variances are summarized, as the Parametric Bootstrap test performed well when dealing with heterogeneous variances. Table 5.5 and Table 5.6 represents the average empirical Type I error rates for $J = 2$ and $J = 3$; respectively under the condition of heterogeneous variances across the three types of distribution.

Table 5.5

Average empirical Type I error rates for J = 2 heterogeneous variances across distributional shapes

Distribution	Test statistics			
	PB test with MAD_n	PB test with T_n	*t*-test	Mann Whitney test
Normal	**0.0267**	**0.0277**	**0.0631**	0.0851
Moderately skewed	**0.0365**	**0.0383**	0.0766	**0.0729**
Extremely skewed	**0.0529**	**0.0535**	0.1166	0.0768

Based on the Table 5.5, the empirical Type I error rates for the two proposed procedures were acceptable with regard to the Bradley's liberal criterion across the three types of distribution. It could be generalized that, Parametric Bootstrap statistic with MAD_n and T_n as trimming criteria are robust under the conditions of heterogeneous variances with normal distribution, moderately skewed distribution and extremely skewed distribution for the two groups case. However, poor results were observed for independent sample *t*-test and Mann Whitney test under extremely skewed distribution with their Type I error rates straying above the 0.075 level. These results indicated that the performance of the proposed procedure with MAD_n

and T_n as new trimming strategy were better than independent sample *t*-test and Mann Whitney test in terms of the ability to control the Type I error for $J = 2$ under extreme condition.

Table 5.6

Average empirical Type I error rates for J = 3 heterogeneous variances across distributional shapes

Distribution	Test statistics			
	PB test with MAD_n	PB test with T_n	ANOVA test	Kruskal Wallis test
Normal	0.0169	0.0187	0.0804	0.0801
Moderately skewed	0.0245	**0.0261**	0.0907	**0.0743**
Extremely skewed	**0.0358**	**0.0441**	0.1229	0.0816

As for the three groups case, the Parametric Bootstrap statistic with MAD_n and T_n as trimming criteria shown in Table 5.6 lose their Type I error control when these procedures were tested on normal distribution under the condition of heterogeneous variances. T_n procedure was found to be robust under moderately skewed distribution and extremely skewed distribution while MAD_n procedure only showed a good control of Type I error rates under extremely skewed distribution. However, the value of the MAD_n procedure was just slightly below the lower interval (0.025) with the average empirical Type I error rates of 0.0245 under moderately skewed distribution. The results from Table 5.6 indicated that only the proposed procedures were proven to be able to control the Type I error under extreme condition while ANOVA test and Kruskal Wallis test were considered not robust with their average empirical Type I error rates for $J = 3$ exceeded the 0.075 level.

Overall, the procedure using the proposed trimming strategy is best used for the condition of skewed distribution and heterogeneous variances compared to classical parametric test (i.e. independent sample *t*-test and ANOVA test) and non-parametric test (i.e. Mann Whitney test and Kruskal Wallis test). It is because Parametric Bootstrap statistic with MAD_n and T_n as trimming criteria were able to produce Type I error rates within the Bradley's liberal criterion of robustness.

Therefore, for $J = 2$, the Parametric Bootstrap statistic is strongly recommended with MAD_n and T_n as new trimming strategy to test the equality of central tendency when dealing with heterogeneous variances with normal distribution and skewed distribution. For the three groups case ($J = 3$), Parametric Bootstrap statistic with automatic trimming criteria, MAD_n and T_n is recommended in testing the central tendency measures under the condition of non-normal distribution and variances heterogeneity.

5.3 Analysis on Real Data

The performance of the Parametric Bootstrap test with MAD_n and T_n was then demonstrated on real data. Two groups were chosen represented normal distribution and equal variances while two groups were chosen represented non-normal distribution and unequal variances. The finding from Table 5.7indicates that the Parametric Bootstrap test with MAD_n and T_n as trimming criteria performed better than the independent sample *t*-test and Mann Whitney test for the case of two groups even for the condition of non-normal distribution and unequal variances.

Table 5.7

p-values for each test

Test statistic	*p*-values	
	Normal data and equal variances	Nonnormal data and unequal variances
PB test with MAD_n	0.1276	0.6951
PB test with T_n	0.0944	0.7172
t-test	0.1516	0.9110
Mann Whitney test	0.1632	0.9240

5.4 Suggestions for Future Research

As stated in the CHAPTER ONE, classical parametric test that are used in this study generally provides a good control of the Type I error only if both assumptions of normality and homogeneity are true. Hence, our main concern is to construct a robust test statistic known as modified Parametric Bootstrap that robust to the violation of these assumptions. This study proved that proposed test statistic performed well under extreme condition which is non-normal distribution and heterogeneous variances even if the sample sizes are unequal by substituting the robust scale estimators, MAD_n and T_n as trimming criteria. However, these two procedures failed to show robustness under homogeneous variances except for T_n procedure when it was tested under the condition of unbalanced sample sizes for the case of two groups.

Therefore, this study should be continued with some other robust scale estimators with a view to find solutions to the conservative Type I error rates since we only focused on two robust scale estimators namely MAD_n and T_n as trimming criteria. Rousseeuw and Croux (1993) in their study suggested plenty of other robust scale estimators that are worth to consider such as Q_n, S_n, and LMS_n. The combination of the Parametric Bootstrap test with MAD_n and T_n showed a good control of the Type I

error rate under extreme conditions. Thus, the Parametric Bootstrap test will compete well with other robust scale estimators.

As for the bootstrap method, Parametric Bootstrap statistic with percentile bootstrap method seems to provide a good control of the Type I error rates in heterogeneous variances cases under skewed distribution. Therefore, percentile Bootstrap test should be adopted with another test statistic.

REFERENCES

Babu, G. J., Padmanabhan, A. R., & Puri, M. L. (1999). Robust one-way ANOVA under possibly non-regular conditions. *Biometrical Journal, 41*(3), 321-339.

Box, G. E. P. (1953).Non-normality and tests on variances. *Biometrika,* 10, 318-335.

Box, G. E. P. (1954). Some theorem on quadratics forms applied in the study of analysis of variance problems: I. Effect of inequality of variance in the one-way model. *Annals of Mathematical Statistics*, 25, 290-302.

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

Cribbie, R. A., Fiksenbaum, L., Keselman, H. J., & Wilcox, R. R. (2012). Effect of non-normality on test statistics for one-way independent groups designs. *British Journal of Mathematical and Statistical Psychology,* 65, 56-73.

Daniel, W. W. (1990). *Applied Nonparametric Statistics*. Boston: PWS-Kent.

Efron, B. (1979). Bootstrap methods: Another look at the Jacknife. *Annals of Statistics*, 7, 1-26.

Efron, B. &Tibshirani, R. (1993). *An Introduction to the Bootstrap*. Chapman & Hall Inc.

Erceg-Hurn, D. M., &Mirosevich, V. M. (2008). Modern robust statistical methods: An easy way to maximize the accuracy and power of your research. *American Psychologist*, 63, 591-601.

Guo, J. H. &Luh, W. M. (2000). An invertible transformation two-sample trimmed *t*-statistics under heterogeneity and nonnormality. *Statistics & Probability letters*, 49, 1-7.

Hampel, F. R. (1974).The influence curve and its role in robust estimation. *Journal of the American Statistical Association,* 69, 383-393.

Hampel, F. R. (2001).*Robust statistics: A brief introduction and overview*. Invited talk in the Symposium "Robust Statistics and Fuzzy Techniques in Geodesy and GIS" held in ETH Zurich, Mar 12-16, 2001.

Hoaglin, D. C. (1985). Summarizing shape numerically: The *g*- and-*h* distributions. In D. Hoaglin, F. Mosteller, and J. Tukey (eds.), *Exploring data tables, trends, and shapes*. New York: Wiley.

Huber, P. J. (1964). Robust estimation of a location parameter. *The Analysis of Mathematical Statistics,* 38, 33-101.

Huber, P. J. (1981). *Robust Statistics*. New York: Wiley.

James, G. S. (1951). The comparison of several groups of observations when the ratios of the population variances are unknown.*Biometrika,* 38, 324-329.

Keselman, H. J., Huberty, C. J., Lix, L. M., Olejnik, S., Cribbie, R. A., Donahue, B., Kowalchuk, R. K., Lowman, L. L., Petoskey, M. D., Keselman, J. C., & Levin, J. R. (1998). Statistical practices of Educational Researchers: An analysis of their ANOVA, MANOVA and ANCOVA analyses. *Review of Educational Research, 68*(3), 350-386.

Keselman, H. J., Kowalchuk, R. K., Algina, J., Lix, L. M., & Wilcox, R. R. (2000). Testing treatment effects in repeated measures designs: Trimmed means and bootstrapping. *British Journal of Mathematical and Statistical Psychology,* 53, 175-191.

Keselman, H. J., Wilcox, R. R., &Lix, L. M. (2003). A generally robust approach to hypothesis testing in independent and correlated group designs. *Psychophysiology*, 40, 586-596.

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

Keselman, H. J., Wilcox, R. R., Othman, A. R., &Fradette, K. (2002). Trimming, transforming statistics, and bootstrapping: circumventing the biasing effects of heteroscedasticity and nonnormality. *Journal of Modern Applied Statistical Methods,* 1, 288-309.

Kohr, R. L. & Games, P. A, (1974). Robustness of the analysis of variance, the Welch procedure, and a Box procedure to heterogenous variances. *Journal of Experimental Education,* 43, 61-69.

Krishnamoorthy, K., Lu, F., & Mathew, T. (2007). A parametric bootstrap approach for ANOVA with unequal variances: Fixed and random models. *Computational Statistics and Data Analysis*, 51, 5731-5742.

Levene, H. (1960). Robust testes for equality of variances. *In Contributions to Probability and Statistics* (I. Olkin, ed.) 278-292. Stanford University Press, CA.

Lix, L. M., & Keselman, H. J. (1998). To trim or not to trim: Tests of location equality under heteroscedasticity and non-normality. *Educational and Psychological Measurement.* Math. Proc. *Cambridge Philosophical Society,* 115, 335-363.

Lix, L. M., Keselman, J. C., & Keselman, H. J. (1996). Consequences of assumption violations revisited: A quantitative review of alternatives to the one-way analysis of variance "*F*" test. *Review of Educational Research,* 66, 579-619.

Manly, B. F. J. (1997). *Randomization, Bootstrap and Monte Carlo Methods in Biology* (*2nd Ed.*). London: Chapman & Hall.

Md Yusof, Z. (2009). Type I error and power rates of robust methods with variable trimmed mean. Unpublished Ph.D. thesis, Universiti Sains Malaysia.

Md Yusof, Z., Abdullah, S., & Syed Yahaya, S. S. (2012a). Type I error rates of parametric, robust and nonparametric methods for two group cases. *World Applied Sciences Journal, 16*(12), 1815-1819.

Md Yusof, Z., Abdullah, S. & Syed Yahaya, S. S. (2012b). Testing the differences of student's scores between two groups. *Journal of Applied Sciences Research*, *8*(9), 4894-4899.

Md Yusof, Z., Harun, N. H., Syed Yahaya, S. S. & Abdullah, S. (2013). A modified parametric bootstrap: an alternative to classical test. *In proceeding of the World Conference on Integration of Knowledge 2013*, Langkawi, Malaysia.

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

Mehta, J. S., & Srinivasan, R. (1970). On the Behrens-Fisher problem. *Biometrika*, 57, 549-655.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletion*, 156-166.

Miller, R. L. & Brewer, J. D. (2003).*The A-Z Social Research*. London: SAGE Publications, Ltd.

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

Othman, A. R., Keselman, H. J., Padmanabhan, A. R., Wilcox, R. R. & Fradette, K. (2003). An improved Welch-James test statistic.*In proceeding of the Regional Conference on Integrating Technology in the Mathematical Sciences 2003*, Universiti Sains Malaysia, Pulau Pinang, Malaysia.

Othman, A. R., Keselman, H. J., Padmanabhan, A. R., Wilcox, R. R. & Fradette, K. (2004). Comparing measures of the 'typical' score across treatment groups. *British Journal of Mathematical and Statistical Psychology*, 215-234.

Othman, A. R., Keselman, H. J., Wilcox, R. R., Fradette, K., & Padmanabhan, A. R. (2002). A test of symetri. *Journal of Modern Applied Statistical Methods*, 1, 310-315.

Rocke, D. M., Downs, G. W., &Rocke, A. J. (1982). Are robust estimator really necessary?. *Technometrics*, *24*(2), 95-101.

Rosenberger, J. L., & Gasko, M. (1983). Comparing location estimators: Trimmed means, medians and trimean. In D. Hoaglin, F. Mosteller& J. Tukey (Eds.). Understanding robust and exploratory data analysis, 297-336. New York: Wiley.

Rousseeuw, P. J., &Croux, C. (1993). Alternatives to the median absolute deviation. *Journal of the American Statistical Association,* 88, 1273-1283.

SAS Institute Inc. (1999): *SAS/IML User's Guide version 8*. Cary. NC: SAS Institute Inc.

Scheffe, H. (1959). *The Analysis of Variance*. New York: Wiley.

Shapiro, S. S. &Wilk, M. B. (1965). An analysis of variance test for normality. *Biometrika*, 52, 591-611.

Staudte, R. G. & Sheather, S. J. (1990). *Robust estimation and testing*. New York: Wiley.

Syed Yahaya, S. S. (2005). Robust statistical procedures for testing the equality of central tendency parameters under skewed distributions.Unpublished Ph.D. thesis, Universiti Sains Malaysia.

Syed Yahaya, S. S., Othman, A. R., &Keselman, H. J. (2006). Comparing the "typical score" across independent groups based on different criteria for trimming. *Methodological Papers, 3*(1), 49-62.

Welch, B. L. (1951). On the comparison of several means: An alternative approach. *Biometrika,* 38, 330-336.

Westfall, P. H. & Young, S. S. (1993). *Resampling-based Multiple Testing*. New York: Wiley.

Wilcox, R. R. (1997). *Introduction to robust estimation and hypothesis testing.* Academic Press, New York.

Wilcox, R. R. (2002). Understanding the practical advantages of modern ANOVA methods. *Journal of Clinical Child and Adolescent Psychology*, 31, 399-412.

Wilcox, R. R. (2005). *Introduction to robust estimation and hypothesis testing* (2nd Ed.). San Diego, CA: Academic Press.

Wilcox, R. R., &Keselman, H. J. (2003). Modern robust data analysis methods: Measures of central tendency. *Psychological Methods,* 8, 254-374.

Wilcox, R. R., & Keselman, H. J. (2010). Modern robust data analysis methods: Measures of central tendency. 1-43.

Wu, M., & Zuo, Y. (2009). Trimmed and Winsorized means based on a scaled deviation. *Journal of Statistical Planning and Inference*, *139*(2), 350-365.

Zikmund, W. G., Babin, B. J., Carr, J. C., & Griffin, M. (2010).*Business research methods* (8th ed.). Thousand Oaks, CA: Thomson/South-Western.

Zimmerman, D. W. (2000). Statistical significance levels of nonparametric tests biased by heterogeneous variances of treatment groups. *Journal of General Psychology,* 127, 354-364.