TESTING THE EQUALITY OF CENTRAL TENDENCY MEASURES USING T_1 STATISTIC WITH DIFFERENT TRIMMING STRATEGIES

SHARIPAH SOAAD SYED YAHAYA
SUHAIDA ABDULLAH
ZAHAYU MD YUSOF

SCHOOL OF QUANTITATIVE SCIENCES
UUM COLLEGE OF ARTS AND SCIENCES
UNIVERSITI UTARA MALAYSIA
2011

ACKNOWLEDGEMENTS

First and foremost, we would like to extend our sincere thanks to Universiti Utara Malaysia (UUM), which funded this research. Our sincere appreciation also goes to our research assistant Miss Norliana Md Yusof, who tirelessly had provided us with all the assistance needed. Finally, our grateful recognition are due to our family, fellow colleagues and friends who had offered guidance and ideas as well as encouragement that has definitely push us to complete this research

Zahayu Md Yusof
Suhaida Abdullah
Sharipah Soaad Syed Yahaya

PENGAKUAN TANGGUNGJAWAB (DISCLAIMER)

Kami, dengan ini, mengaku bertanggungjawab di atas ketepatan semua pandangan, komen teknikal, laporan fakta, data, gambarajah, ilustrasi, dan gambar foto yang telah diutarakan di dalam laporan ini. Kami bertanggungjawab sepenuhnya bahawa bahan yang diserahkan ini telah disemak dari aspek hakcipta dan hak keempunyaan. Universiti Utara Malaysia tidak bertanggungan terhadap ketepatan mana-mana komen, laporan, dan maklumat teknikal dan fakta lain, dan terhadap tuntutan hakcipta dan juga hak keempunyaan.

We are responsible for the accuracy of all opinion, technical comment, factual report, data, figures, illustrations and photographs in the article. We bear full responsibility for the checking whether material submitted is subject to copyright or ownership rights. UUM does not accept any liability for the accuracy of such comment, report and other technical and factual information and the copyright or ownership rights claims.

Ketua Penyelidik:

____
Tandatangan

Nama: Dr. Zahayu Binti Md Yusof

Ahli:

____
Tandatangan

Nama: Pn. Suhaida Abdullah

____
Tandatangan

Nama: Prof. Madya Dr. Sharipah Soaad Syed Yahaya

TABLES OF CONTENTS

CHAPTER FOUR: RESULTS OF THE ANALYSIS

CHAPTER FIVE: DISCUSSION AND CONCLUSION

LIST OF TABLES

LIST OF FIGURES

Page

LIST OF ABBREVIATIONS

ANOVA	Analysis of variance
HQ1	A hinge estimator
LMS_n	A scale estimator
MAD_n	Median absolute deviation about the median
T_1	A statistical method for testing the equality of central tendency measures
T_n	A scale estimator

LIST OF PUBLICATIONS

Md. Yusof, Z., Abdullah, S., Syed Yahaya, S. S. & Othman, A. R. (2011). (2011).Type I error rates of F_t statistic with different trimming strategies for two groups case. Modern Applied Science, **5(4)**, 236 – 242.

Md. Yusof, Z., Abdullah, S., Syed Yahaya, S. S., & Othman, A. R. (2011).Testing the equality of central tendency measures using varies trimming strategies. *African Journal of Mathematics Computer Science Research*, **4(1)**, 32-38.

Md Yusof, Z., Abdullah, S., & Syed Yahaya, S. S. (2010). Type I error rates of *t*-test and T_1 statistic under balanced and unbalanced designs. *Prosiding Seminar Kebangsaan Sains dan Matematik 2010*. Kompleks Pentadbiran Kerajaan Persekutuan, Kota Kinabalu Sabah.

TESTING THE EQUALITY OF CENTRAL TENDENCY MEASURES USING T_1 STATISTIC WITH DIFFERENT TRIMMING STRATEGIES

ABSTRACT

When the assumptions of normality and homoscedasticity are met, researchers should have no doubt in using classical test such as *t*-test and *ANOVA* to test for the equality of central tendency measures for two and more than two groups respectively. However, in real life we do not often encounter with this perfect situation. T_1 statistic was proposed as an alternative robust method that could handle the problem of nonnormality when using trimmed mean with 15% symmetric trimming as the central tendency measures, but their study only focused on the condition of homogeneous variances. Motivated by the good performance of the method, in this study we propose using T_1 statistic with three different trimming strategies, namely, i) predetermined 15% symmetric trimming ii) predetermined asymmetric trimming based upon hinge estimators and iii) empirically determined asymmetric trimming based on robust scale estimators, MAD_n, T_n *and* LMS_n to handle simultaneously the problem of nonnormality and heteroscedasticity. To test for the robustness of the procedures towards the violation of the assumptions, several variables will be manipulated. The variables are types of distributions, heterogeneity of variances, sample sizes, nature of pairings of group sample sizes and group variances, and number of groups. Type I error for each procedures will then be calculated. This study will be based on simulated data with each procedure will be simulated 5000 times and each set of data will be bootstrapped 599 times. The proposed procedures, generally, generated good Type I error control. The combination of T_1 statistic with HQ_1 produced promising procedures that are capable of addressing the problem of testing the equality of central tendency measures especially for skewed distributions.

Keywords: Robust statistics, Type I error, robust scale estimators, skewed distributions

CHAPTER 1

INTRODUCTION

1.1 Introduction

In recent years, numerous methods for locating treatment effects or testing the equality of central tendency (location) parameters by simultaneously controlling the Type I error to detect effects are being studied. Progress has been made in terms of finding better methods for controlling the Type I error of the test that detects treatment in one-way independent group designs (Babu *et al*., 1999; Othman *et al*., 2004; Wilcox and Keselman, 2003). Through a combination of impressive theoretical developments, more flexible statistical methods, and faster computers, serious practical problems that seemed insurmountable only a few years ago can now be addressed. These developments are important to applied researchers because they greatly enhance the ability to discover true differences between groups while maximizing the chance of detecting a genuine positive effect.

The parametric approach in testing the equality of the central tendency parameters continued to play a prominent role because of its capacity to comprehensively describe information contained in a data. However, the good performance and valid application of the procedures require strict adherence to certain assumptions, which do not always operate as predicatively as assumed in the real world. Some of the most common statistical procedures are extremely sensitive to these minor deviations from assumptions such as in the case of normality of distributions and homogeneity of variances**.** As an example, when computing confidence intervals and testing hypothesis about means, the methods are based on the assumption that observations are randomly sampled from normal distributions. Another instance is when comparing independent groups; where the methods are also assume that groups have a common variance. Currently, these methods form the backbone of most applied research that involves statistical methodology. It is therefore desirable to construct methods of inference that do not depend on distributional and homoscedasticity (equal variances) assumptions for their validity.

Consequently, nonparametric statistics emerged as a field of research and some of its methods become widely popular in applications. The basic principle was to make as few assumptions about the data as possible and still get the answer to a specific question. However, nonparametric procedures are more appropriate for data based on weak measurement scales. Besides, procedures in the nonparametric are less powerful than the parametric and therefore, require a larger sample size to reject a false hypothesis. In practice, it often happens that we need to robustly estimate central tendency and/or scale from small sample. The sample size *n* is often constrained by the cost of an observation. In many experimental settings (e.g. in chemistry) one will typically repeat each measurement only a few times. Even a small sample may contain aberrant values due to technical problems or measurement inaccuracies for example, and since the sample is small, getting rid off the aberrant values is very much avoidable.

1.2 Robust Statistics

There are several definitions of robust statistics that have been found in the literature and these unfortunately lead to the inconsistency of its meaning. Most of the definitions are based on the objective of the particular study by different researchers (Huber, 1981).

Robust statistics combine the virtues of both, the parametric and the nonparametric approach**.** In nonparametric inference, few assumptions are made regarding the distribution from which the observations are drawn. In contrast, the approach in robust inference is different wherein there is a working assumption about the form of the distribution, but we are not entirely convinced that the assumption is true. Robustness theories can be viewed as stability theories of statistical inference. What is desired is an inference procedure, which in some sense does almost as well as possible if the assumption is true, but does not perform much worse within a range of alternatives to the assumption.

A statistical method is considered robust if the inferences are not seriously invalidated by the violation of such assumptions, for instance nonnormality and variance heterogeneity (Scheffe, 1959). Huber (1981) defined robustness as a situation which is not sensitive to small changes in assumptions while Brownlee (1965) reported slight effects on a procedure when appreciable departures from the assumptions were observed.

The theory of robust statistics deals with deviations from the assumptions on the model and is concerned with the construction of statistical procedures which is still reliable and reasonably efficient in a neighborhood of the model (Ronchetti, 2006). Hampel *et al.* (1986), stated that in a broad informal sense, robust statistics is a body of knowledge, partly formalized into "theories of robustness" relating to deviations from idealized assumptions in statistics. As mentioned by Hoel *et al.* (1971), a test that is reliable under rather strong modifications of the assumptions on which it was based is said to be robust. Hence in this research, a statistical method is considered robust when it has estimators which cannot be influenced by the deviations from the given assumptions when hypothesis testing is being conducted.

The classical tests of group equality such as the *t* test and analysis of variance (ANOVA) are always misrepresented due to variance heterogeneity and nonnormality. To overcome the problem of variance heterogeneity, many methods were developed and proved to be more robust than these classical tests. A few of such methods are the Welch test (Welch, 1951), the James test (James, 1951) and the Alexander-Govern test (Alexander & Govern, 1994). These methods are among the best methods that have good control of type I error rates (Alexander & Govern, 1994; Schneider & Penfield, 1997; Myers, 1998).

In their efforts to control the Type I error rate, investigators looked into numerous robust methods since these methods generally are insensitive to assumptions about the overall nature of the data (e.g. Babu *et al*., 1999; Keselman *et al*., 2004b; Kulinskaya, 2003; Luh and Guo, 1999; Othman *et al*., 2004). Any small deviations from the model assumptions should only slightly impair the performance, for example, the level of a test should be close to the nominal value calculated at the model, and larger deviations from the model should not cause catastrophe**.** Robust measures of central tendency such as trimmed means, medians or *M*-estimators (refer to Huber, 1981; Staudte and Sheather, 1990; Wilcox, 2005) have been considered as alternatives for the usual least squares estimator, i.e., the usual least squares means, in most research recently (e.g. Keselman *et al*., 2004; Luh and Guo, 1999; Wilcox *et al*., 1998; Wilcox and Keselman, 2002). These measures of central tendency had been shown to have better control over Type I error and power to detect treatment effects (see e.g. Lix and Keselman, 1998; Othman *et al*., 2004; Wilcox, 2005; Yuen, 1974). Yuen (1974) found these benefits in the two-group case of trimmed means and Lix and Keselman (1998) demonstrated similar results in the more than two-group problem. Other investigators, e.g. Babu, *et al*. (1999) used median as the central tendency measure when dealing with skewed distribution and Wilcox and Keselman (2003) introduced a modified one-step *M*-estimator (*MOM*) as the central tendency measure when testing for treatment effects.

To date, there are several new procedures that were developed to deal with group trimmed means. Even though the usual trimmed mean has good control of Type I error rate, the trimming is done regardless of the types of distribution. The percentage of trimming is set prior to the facts whether the outliers are presence or not. It will be a gross mistake to eliminate data which are not outliers such as in a normally distributed data. Keselman *et al*. (2007) proposed an adaptive trimmed mean which trimmed extreme data based on the types of distribution as an alternative to the usual trimmed mean. This adaptive trimmed mean uses hinge estimator HQ_1 (Reed & Stark, 2003) in order to adjust the trimming process that suits the shape of data distribution. Keselman *et al.* (2007) successively improved Welch test using this adaptive trimmed mean in controlling Type I error rates.

Another method which uses trimmed mean is the modified *MOM-H* statistic introduced by Wilcox and Keselman (2003) which used modified one-step *M*-estimator *(MOM)* as the central tendency measure in their work on the *H* statistic. Essentially, *MOM* is automatic variable trimming. This method was proven to have good control of Type I error rates when comparing for the differences between distributions. Motivated by the good performance of this procedures, in this research we propose a modification of T_1 statistic developed by Babu *et al*. (1999) with three different trimming strategies namely i) predetermined 15% symmetric trimming ii) predetermined asymmetric trimming based upon hinge estimators (Keselman, Wilcox, Lix, Algina & Fradette, 2007) and iii) empirically determined asymmetric trimming based on robust scale estimators, *MAD_n, T_n and LMS_n* (Rousseeuw & Croux, 1993) to handle simultaneously the problem of nonnormality and heteroscedasticity.

1.3 Trimming

Two approaches that may be considered by researchers faced with data that appear to violate the ANOVA assumptions are (i) to apply a transformation to the data and proceed with use of the *F* test or (ii) to select an alternative test procedure which is insensitive (i.e., robust) to assumption violations.

1.3.1 Purpose of trimming

When data are not normal and variances are heterogeneous, it is often possible to transform the data so that the new scores more nearly approximate normality and equality of variances. For example, when dealing with skewed distributions, two general suggestions are to take the square root or logarithms of every observation. Often these transformations produce data that are nearly normal. In some circumstances, the same transformations also achieve equality of variances (Maxwell & Delaney, 2004). Transforming data from designed experiments is an old and valuable tool (Carroll, 1982). Most researchers would wish to transform data if such was necessary to obtain a normal distribution. Upon transformation, standard analyses will often be performed.

However, there are some issues that should be kept in mind when applying transformation. First, when doing transformation on the data, it indicates that an attempt at making inferences about the mean of the original score has been ignored. This will lead to complex issues of interpretation, since the conclusions which are drawn must be based on the transformed scores, not the original observations (Lix *et al.*, 1996). Thus, the interpretation of the results may also be less clear (Maxwell & Delaney, 2004). For example, most individuals find it difficult to understand the mean value of the square root of their original observations. Second, the complex transformations (i.e. Box-Cox transformation) do not remove the effects of outliers. That is, outliers remain and can inflate the sample variance and also lower the power by a substantial amount. Third, if each observation is transformed in the same manner, situations arise where the distribution of the observed scores remains skewed (Wilcox, 2002). Fourth, there is the problem of finding the correct transformation. Even though, there are a variety of transformations which may be applied to a set of data (Oshima & Algina, 1992), depending on the particular type and degree of assumption violation that is thought to be present in the data, this may not always be a simple solution (Lix *et al.*, 1996). Also, it is difficult to find a transformation that will simultaneously deal with asymmetric and variance heterogeneity (Keselman *et al*., 2007).

Because of all of these drawbacks especially the interpretation issues, e.g. square root of the mean and log of the mean, we will ignore transformation and consider a robust method involving trimming.

Robust method is another alternative method to deal with nonnormal distribution. A robust test will control the actual Type I error rate close to the nominal level of significance, even when the data do not conform to the test's derivational assumptions, and will maintain actual statistical power close to theoretical power, as well (Lix *et al.*, 1996). The literature so far suggest that the robust test are generally superior to the *F* test in the majority of assumption violation situations where the classical ANOVA *F*-test and alternative test statistics (e.g., Welch) would not be (e.g., Levy, 1978; Tomarken & Serlin, 1986).

Methodology researchers consider ways to improve the performance of alternative procedures when the data are nonnormal (Lix *et al.*, 1996). Wilcox (1995) has suggested that trimming, or discarding outliers from a data set prior to analysis, can lead to improve performance, both in terms of Type I error control and power. Trimming is the most popular robust based method when dealing with skewed data. Naturally, trimming is a very drastic way of dealing with extreme observations. However, removing a small set of observations in a relatively large sample should not change the results in a major way (Rodrigues & Rubia, 2006).

The key factors in trimming are the amount of trimming and how the trimming is specifically conducted. There are two common methods in trimming, symmetric and asymmetric trimming. In symmetric trimming, equal amount of trimming is applied on both tails of the distribution. In asymmetric trimming, the process of trimming is either conducted on one-tail or on both tails with unequal amounts. In order to avoid loss of information, trimming need to be conducted with care. Before trimming could be performed, the amount of trimming has to be determined first, usually by fixing the amount of trimming (predetermined). In our study, we are going to depart from trimming with fixed amount to automated trimming.

1.3.2 Trimmed mean

Trimming will definitely get rid of outliers but how do we address the question of outliers? Usually outliers are culprits leading to nonnormality and heterogeneity.

Even so, if we are looking at the differences between groups, the presence of outliers in one group will definitely lead to rejection of the null hypothesis. How do we deal with this rejection? The rejection will not be taken at face value. Further analysis will now be done on the outliers. However, in our study, the question of outliers does not arise because our study conditions do not involve them. Our study conditions are variance heterogeneity, pairing of group variances and group sample sizes, types of distributions, balanced and unbalanced sample sizes and number of groups.

Trimmed mean is a central tendency measure that summarizes data when trimming is carried out. By using the trimmed means, the effect of the tails of the distribution is reduced by their removal based on the trimming percentage that has to be stated in advanced (predetermined amount). The common trimmed mean used the fixed amount of trimming method. It needs the fix amount of trimming percentage and tight down with this amount of trimming. By using this method, amounts such as 10% or 20% of the observations from a distribution will be trimmed from both tails. In the case of a light-tailed distribution or the normal distribution, it may be desirable to trim a few observations or none at all. There is extensive literature regarding this trimming method that uses the fixed amount of symmetric trimming. Among them are Lee and Fung (1985), Keselman *et al.* (2002), and Wilcox (2003).

If we have skewed distributions then the amounts of trimming on both tails should be different. More should be trimmed from the skewed tail. However, if the fixed symmetric trimming is used, regardless of the shape of the tails, the trimming is done symmetrically as set. A research by Keselman *et al*. (2007) used asymmetric trimming and in particular, applying hinge estimators proposed by Reed and Stark (1996) to determine the suitable amount of trimming on each tail of a distribution. However, their method still used fixed trimming percentages.

The trimmed mean is not so robust because the breakdown point of trimmed mean is just as much as the percentage of trimming and this shows that trimmed mean cannot withstand large numbers of extreme value. Wilcox *et al.* (2000) in their study stated that when comparing trimmed means versus means with actual data, the power of the trimmed mean procedure was observed to be greatly increased. They also discovered that there was improved control over the probability of a Type I error.

1.4 T_1 Statistic

Types of distributions and homogeneity of variances are two important aspects that need to be taken into consideration before we proceed with the testing of the equality of central tendency measures using robust statistics. If the type of distribution is unknown and cannot be assumed as normally distributed, Babu *et al*. (1999) suggested the use of their T_1 statistic to compare the differences between distributions. They applied this statistic when the distributions are tested symmetric. This procedure used 15% symmetric trimming with trimmed mean as the central tendency measure.

1.5 Objective of the Study

The main objective of this study is to examine the operating conditions that would result in good Type I error rates for the following new procedures:

1. T_1 with predetermined 15% symmetric trimming.
2. T_1 with predetermined asymmetric trimming based upon hinge estimators (Keselman, Wilcox, Lix, Algina & Fradette, 2007).

3. T_1 with automatic trimming based on robust scale estimators, MAD_n, T_n and LMS_n (Rousseeuw & Croux, 1993)

The secondary objective is to compare the performance of procedures 1 – 3 via Type I error. In doing so, this study should be able to

1. determine the best trimming strategy.
2. recommend the best procedure for extreme conditions.

1.6 Significance of the Study

This research will contribute towards knowledge development in experimental design methodology especially in the experimental sciences. Statisticians are aware that experimental design methodology depends on assumptions of normality and treatment groups having equal variances. However, in the real world, data are not always normally distributed. The benefit of this research is that with these new alternative methods, researchers (in various fields, especially the experimental sciences) will not be constrained with all the assumptions such as normality and homogeneity of variances. They can instead work with the original data without having to worry about the shape of the distributions. This research contributes to the development of robust statistics that uses trimming strategy in its test statistic or in its procedures. Robust statistics with trimming strategy were designed to handle assumptions of normality and variance homogeneity. This research will also naturally want to determine which trimming strategy is the best for the T_1 statistics.

1.7 Organization of the report

Chapter 1 gives an introduction on the importance of the study and gives in depth explanation regarding the robust statistical methods. This chapter also presents a brief introduction to the methods proposed in this study, namely T_1 statistics. Details of these methods are presented in Chapter 2. Chapter 2 also discusses about the scale estimators and defines terminologies used throughout this study. Explanations about operating conditions that have been manipulated are found in Chapter 3. They are the number of groups, the sample sizes for balanced and unbalanced design, heterogeneity of variances, the nature of pairings of group sample sizes and group variances and type of distributions. This chapter further gives the design specifications and explains the generation of data used in this study. The results from the analyses of Type I error was presented in Chapter 4. We conclude our findings and propose suggestions for further studies in the last chapter.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

The two sample *t*-test and the analysis of variance (ANOVA) are two common statistical methods used to locate treatment effects in a one-way independent group design. However, in using these two statistics, assumptions of normality and variance homogeneity need to be fulfilled. In real life applications, these conditions are rarely achieved and these will lead to inaccuracy in decision based on the testing procedure.

Departures from normality originate from two problems, i.e. skewness and the existence of outliers. These problems could be remedied by using transformation such as exponential, logarithm and others but sometimes, even after the transformation, problems with nonnormal data still occur. Simple transformations of the data such as by taking logarithm can reduce skewness but not for complex transformations such as the class of Box-Cox transformations (Wilcox & Keselman, 2003). However, problems due to the outliers are not eliminated. According to Wilcox and Keselman (2003), a simple transformation can alter skewed distributions to make them more symmetrical, but they still do not deal directly with outliers. They suggested using a trimming method when dealing directly with outliers.

The existence of outliers in a sample data will cause the probability of Type I error to be less than the nominal alpha level and concurrently lower the power of the test statistic. In the application of *t*-test, outliers can inflate the sample variance and simultaneously lower the value of the test (Wilcox & Keselman, 2003). Even when sampling from a perfectly symmetrical distribution, outliers can still cause the *t*-test to lose power when compared against modern methods. Modern methods here are methods that are based on robust measures of location (Wilcox & Keselman, 2003). According to Keselman, Lix *et al.* (1998), the reduction in the power to detect differences between groups occurs because the usual population standard deviation is greatly influenced by the presence of the extreme observations in a distribution of scores.

The presence of outliers will inevitably lead to the observed scores being skewed. However, skewness itself can be an inherent property of several score distributions. It is also well known that skewness can also be a problem when we are trying to control the probability of Type I error. Type I error rates and the confidence intervals can be highly inaccurate when the data are skewed. For the normal distribution and any symmetric distribution, the skewness for the distributions are zero. When the data are skewed to the left, the skewness value is negative. This denotes that the left tail is longer than the right tail. When the data are skewed to the right, the skewness value will be positive. Many classical statistical tests depend on normality assumptions. When this assumption is not satisfied, the rate of Type I error and the power of the test conducted will be affected.

The sample mean is the most common estimator used in most statistical analyses. However, this estimator is very sensitive to the presence of outliers and skewness. One single outlier could easily influence this estimator, thus causing it to have a low breakdown point (Sawilowsky, 2002). In addition, the sample mean also has unbounded influence function, implying that a single contaminated observation may have a considerable effect on the estimate (Thomas, 2000). Under these conditions, any test that used the sample mean as the estimator will produce low power and distorted rates of Type I error. These include the *t*-test and ANOVA. Furthermore, the standard error of the usual mean can become seriously inflated when the underlying distribution is heavy-tailed. To address this problem, Wilcox and Keselman (2003) suggested using estimators of robust measures of location and rank-based methods. Some of these robust estimators are the *M*-estimator and trimmed mean.

The sample trimmed mean (will be referred to as “trimmed mean” throughout this thesis) is one of the estimators which are able to handle the problem of nonnormality due to skewness. When using this estimator, the smallest and the largest observations in the distribution will be trimmed, thus automatically discarding skewed data. By using the trimmed mean, high power, accurate probability coverage, relatively low standard errors, a negligible amount of bias and a good control over the probability of a Type I error can be achieved (Wilcox & Keselman, 2003).

There are two possibilities of estimating the trimmed mean, i.e. equal amount of trimming or symmetric trimming and unequal amount of trimming or asymmetric trimming. In symmetric trimming, the trimming is done equally on both sides of the distribution. While for asymmetric trimming, the trimming is done on only one side or unequally on both sides of the distribution. Othman *et al.* (2002) in their study suggested that when the data are said to be skewed to the right, then in order to achieve robustness to nonnormality and greater sensitivity to detect effects, one should trim data just from the upper tail of the data distribution. Hogg (1974), Hertsgaard (1979), and Tiku (1980, 1982) suggested that the data should have different amounts of trimming percentages from the right and left tails of the distribution. Keselman *et al*. (2007) proposed a method called adaptive robust estimators to determine the number of observations to be trimmed from each tail of the distribution. By using this method, the total amount of trimming is determined a priori before making the decision whether to trim the data symmerically, asymmetrically or not to trim at all.

If the distribution is skewed, the trimmed mean provides better estimates of the typical score than the usual mean. This is due to the fact that when a distribution is skewed, the trimmed mean does not estimate μ but rather some value (i.e. μ_t) that is typically closer to the bulk of the observations (Keselman *et al.*, 2004). Herron and Hillis (2000) stated that, for heavy-tailed distributions, the trimmed mean is less sensitive to the outliers and also have smaller standard errors than the usual mean. To avoid unnecessary loss of information due to trimming, if a distribution is highly skewed to the left, it seems more reasonable to trim more observations from the left tail of the distribution than from the right tail.

However, the trimmed mean suffers from at least two practical concerns which are (i) the proportion of data at the tails exceeds the percentage of adopted trimming and vice versa and (ii) the trimming is done unproportionately. In the latter case, the problem occurs when equal percentage of trimming (as in trimmed mean) on both tails is adopted on skewed distribution, whereas it would be more reasonable to trim more observations from the tail that is highly skewed. Note that these problems arise because of the amount of trimming have to be fixed in advance without examining the characteristics of the data. In many situations, researchers would want to use an adaptive trimmed mean, (i.e. asymmetric trimmed mean) in which the trimming proportion adapts itself to the characteristics of the distribution on the basis of the sample.

To avoid from trimming erroneously, the process needs to be done meticulously. In our proposed method of trimming, this problem can be avoided since the amount of trimming is determined by the characteristics of the sample data. This method utilizes characteristics of the observed data to determine whether data should be trimmed symmetrically, asymmetrically or not at all. The idea is that, good efficiency will be obtained when sampling from normal distributions as well as non-normal distributions by introducing flexibility into how much is trimmed.

Another problem which researchers always encountered when using the classical methods is heteroscedasticity. Some of the parametric methods that can handle this problem are those proposed by Welch (1961), James (1951) and Alexander and Govern (1994). Unfortunately, all of these methods have difficulty in dealing with problem of nonnormal data. Nonetheless, Abdullah *et al.* (2008) found that Alexander and Govern test which uses automatically trimmed mean as the central tendency measure in place of the usual mean is robust to skewed data when the trimming strategy was adopted.

Some researchers sought for alternatives in the non-parametric methods, such as Mann Whitney and Kruskall Wallis. However, these methods have low power (Wilcox, 1992). Even though non-parametric methods are distribution free, they are not assumptions free. Usually the distribution has to be symmetric. The alternative is to use a robust approach to deal with the problems of nonnormality and heteroscedasticity.

Robust statistics combine the virtues of both, the parametric and the non-parametric approach. In general, these statistics are used in handling the problem of the violation of the independence assumptions such as nonnormality and variance heterogeneity. In this study, we suggested robust procedure, the T_1 statistic proposed by Babu *et al*. (1999). Babu *et al*. (1999) suggested the use of T_1 statistic to compare the differences between distributions if the type of distribution is unknown and cannot be assumed as normally distributed. They applied this statistic with 15% symmetric trimmed mean as the central tendency measure when the distributions are tested symmetric. Trimmed *F* statistic is a statistical method that is able to handle problems with sample locations when nonnormality occurs but the homogeneity of variances assumption still applies.

In this study, we will look at the problems of nonnormality and variance heterogeneity, simultaneously. We will use these statistics with trimming strategies using robust scale estimators, T_n and LMS_n proposed by Rousseeuw and Croux (1993). In addition to these two estimators, we also consider one of the most popular estimators, MAD_n. We choose these estimators because of their high breakdown points and bounded influence functions. These strategies will trim extreme values without the need to state the trimming percentage in advanced.

There are a few terminologies that will be used throughout our study. We will discuss these terminologies briefly in the next sections prior to the in depth discussion of the proposed methods.

2.2 Trimming

Trimming is a method to eliminate outliers or extreme observations from each tail of a distribution. Determining the percentage of trimming must be made prior to the testing. In order to make this decision, efficiency is one factor to be considered. In this context, efficiency means achieving relatively small standard error when the trimming method is used. Trimming needs to be done cautiously. If the amount of trimming is too small, efficiency can be very poor when sampling is from heavy-tailed distribution, but if the amount is too large, efficiency will be very poor when we consider the sampling from a normal distribution (Keselman *et al.*, 2000).

Trimming can be very beneficial in terms of efficiency and in achieving high power. Trimming can eliminate outliers and power might be increased substantially. This is a conclusion that follows almost immediately from a result derived by Laplace two centuries ago (Wilcox, 2005b). According to Wilcox (1998) trimming can be good or bad in terms of power, depending upon the criteria we adopt and the goals we hope to achieve. In Wilcox (2005b), it is stated that the median corresponds to the most extreme case in which all but one or two values are trimmed. He gave an example that if *n* is even, all but two observations are trimmed and if *n* is odd, all but one. Due to the extreme amount of trimming reflected by the usual sample median, the sample median will have a large standard error and low power relative to using the usual sample mean (Wilcox, 2005b).

Theory indicates that the more we trim, the more we can reduce problems due to skewness. Rocke *et al.* (1982) in their paper concluded that the best results were obtained with 20% – 25% symmetric trimming, while Othman *et al*. (2004) reported that one can achieve a slightly better Type I error control with a 15% symmetric trimming rather than a 20% symmetric trimming. Keselman, Othman *et al*. (2004) demonstrated that good control of Type I error can be achieved with only modest amounts of trimming, namely 15% or 10% from each tail of the distribution. For long-tailed symmetric distributions, Lee and Fung (1985) recommended the used of 15% symmetric trimming. According to the literature, the optimal fixed amount of symmetric trimming percentage is between 0% and 25%.

When sampling from a symmetric distribution, it is intuitively appealing to use symmetric trimming (Wilcox, 2003). Symmetric trimming trims the same number of observations at both ends of data and hence is quite efficient for symmetric distributions. However, this strategy becomes less efficient when there is even just a slight departure from symmetry, for example with one end containing outlying points (Wu & Zuo, 2009). Higher amount (i.e. more than 20%) of symmetric trimming should be used when sampling from a skewed distribution (Wilcox, 2003). Nevertheless if the amount of trimming is too high, this can result in lower power when sampling from a light tailed distribution (i.e. normal distribution) where outliers are relatively rare. While for heavy-tailed distributions, the power goes up as the amount of trimming increases, (Wilcox, 1995).

It has been a general practice that 90%, 95%, and 99% are typical choices to specify coverage probabilities. Nevertheless, as stated in Granger (1996), practical forecasters seem to prefer 50% intervals whereas academic writers focus almost exclusively on 95% intervals. It is noted that the larger the probability coverage, the wider the prediction interval, and vice versa. Relating to the trimming percentages, Wilcox (1998) stated that the more we trim, the less effect skewness had on the probability coverage. According to Wilcox (1996), a 20% trimming provide more accurate probability coverage of confidence intervals regarding differences between means when the distributions are skewed.

Nevertheless, when the sample size, *n* is small, the optimal amount of trimming is yet to be determined. The amount of trimming can also be arrived at empirically. However, it is difficult to do so. This is usually attempted when doing one-sided or asymmetric trimming. Othman *et al*. (2002) dealt with predetermined amount of trimming on one side. The recent study done by Keselman *et al.* (2007) also worked with fixed total amount of trimming for both sides of the distribution. They then identified the number of observations that should be trimmed from each tail by the characteristics of the sample data. However, the total number of trimmed data from the left and right tail of the distribution must be equal to the total amount of trimming that they determined earlier. The mismatch of the proportion of skewed data is still of practical concern if we use this method. Thus, in this study, we proposed a method of trimming without any fixed amount. The amount of trimming for both tails of the distribution is determined automatically using robust scale estimators, namely, MAD_n, T_n and LMS_n to get the sample values. We also compared this automatic method of trimming with the usual symmetric trimming. Specifically we chose 15% symmetric trimming for this purpose.

Essentially one does not trim a fixed amount of the data but only the skewed data. These trimming mechanisms will ensure that the problems of outliers and skewed data will be adequately addressed.

2.3 Type I Error

Hypothesis testing is the art of testing if variation between sample distributions can either be explained by chance or not. If we are to test two distributions to see if they vary in a meaningful way, we must be aware that the difference is not just by chance. Type I error is the error of rejecting the null hypothesis given that it is actually true. In other words, this is the error of accepting an alternative hypothesis when the results can be attributed to chance.

According to Steven (1990), a test statistic is robust if the actual level of significance is very close to the nominal level. The nominal level is the level set by the experimenter and is the percent of time one rejects falsely when the null hypothesis is true and all assumptions are met. While the actual level is the percent of time one rejects falsely if one or more of the assumptions are violated.

Type I error rejects an idea that should not have been rejected and also claims that two observations are different, when they are actually the same. It is also known as a ‘false positive’. A false positive usually means that a test claims something to be positive, when that is not the situation. The probability of a Type I error is designated by the Greek letter alpha (α) and is called the Type I error rate.

Conventionally Type I error is set at 0.05 or 0.01. This brings the meaning of there is only 5 or 1 in 100 chance that the variation that we obtained is due to chance. This is called the 'level of significance'. The significance levels need to be chosen attentively. For example, a 5% significance level is the rate to declare a result to be significant when there is actually no relationship in the population. The 5% value is also known as the rate of false alarms or false positives.

By convention, a procedure can be considered robust if it’s Type I error is between 0.5α and 1.5α (Bradley, 1978). Thus, when the nominal level is set at α = 0.05, the Type I error rate should be in between 0.025 and 0.075. Type I error rates are considered liberal when they are above the 0.075 limit while those below the 0.025 limit are considered conservative. However, Guo and Luh (2000) in their study regarded a test with 5% level of significance to be robust if its empirical Type I error rate does not exceed the 0.075 limit.

2.4 Central Tendency Measures

Measures of central tendency are measures of the location of the center of a distribution. The word "center" is purposely left somewhat vague so that the term "central tendency" can refer to a wide variety of measures. It is can also be defined as a single value that summarizes a set of data (Mason, Lind & Marchal, 1999). A measure of central tendency gives the center of a histogram or a frequency distribution curve (Mann, 2004). Some of the examples of central tendency measures are mean, median, mode, trimmed mean and Winsorized mean.

Mean is the most commonly used measure of central tendency. For symmetric distributions, these measures are all the same but for skewed distributions, they can differ markedly. Wilcox (1998) stated that, a very small shift away from normality can inflate σ result in extremely poor power for any hypothesis-testing based on means. A small shift from a normal distribution towards a heavy-tailed distribution will result in large standard errors for the sample mean and in many situations; standard hypothesis-testing methods for means can miss true differences because of low power due to sampling from heavy-tailed distribution (Wilcox, 1998).

In this study, the central tendency measure used is the trimmed mean. By using this measure in place of the usual mean, tests that are insensitive to the combined effects of nonnormality and variance heterogeneity can be obtained. Besides that, Kulinskaya and Dollinger (2007) recommended the use of the trimmed mean because of its simplicity and high relative efficiency.

2.4.1 Trimmed Mean

Trimmed mean is the arithmetic average of residual data after deleting the k smallest observations and the k largest observations. The idea of a trimmed mean is to eliminate *outliers*, or extreme observations. γ% trimmed means were calculated with γ% trimming on both tails. Therefore γ% trimmed means has actually (100 - 2γ) % of data left. The γ% trimmed means is the average of the observations that remain after a proportion γ% has been trimmed from each end of the ordered sample.

How much trimming should be done in a given situation is still questionable, but the important matter is that some trimming often gives substantially better results, as compared to no trimming (Wilcox, 1996). The median is the 50% trimmed mean. This is because when trimming is done on both tails, all the observations except for the median will be eliminated. On the other hand, the arithmetic mean is the 0% trimmed mean. A trimmed mean is apparently less liable to the effects of extreme data than the arithmetic mean. It is then, less liable to sampling fluctuation than the mean for extremely skewed distributions. Trimmed mean is best suited for large data, inconsistent deviations or extremely skewed distributions.

The trimmed mean is a robust estimator of location because it is relatively insensitive to outliers. When the distribution deviates substantially from normality, trimmed mean lessen the effects of outliers and reduce the variance of the estimator over the usual mean (Reed, 1998). Trimmed mean as opposed to the usual least squares means, provide better estimates of the typical individual score in a distribution that contains outliers or is skewed in shape (Keselman, Lix *et al.*, 1998). The standard error of the trimmed mean is less affected by departures from normality than the usual mean because of the censored or the removal of the extreme observations (Keselman, Kowalchuk *et al.*, 2000). Another advantage of trimmed estimators is that, the estimators do not require intensive computations and are intuitively easy to understand (Lix *et al*., 2003). Reed (1996) described the trimmed mean as a safe estimator because its performance does not vary markedly from situation to situation.

2.5 Scale Measures

The central tendency measures do not reveal the whole picture of the dispersion of the distribution of a data set (Mann, 2004). The measures that acknowledge the spread of the data are called scale measures. Three measures that are commonly used are range, variance and standard deviation. Standard deviation is a scale measure with zero breakdown point. So, the standard deviation is not robust. In order to get tests that are not sensitive to the effects of nonnormality and variance heterogeneity, the Winsorized variance is adopted in this study.

2.5.1 Winsorized Variance

In the quest to get the Winsorized variance, the trimmed mean must be calculated first. The Winsorized variance is a consistent estimator of the variance of the corresponding trimmed means (Gross, 1976). The finite sample breakdown point of Winsorized variance denoted as s_w^2 is γ and this is the reason why the standard error of the trimmed mean is less affected by heavy-tailed distribution as compared to the mean (Wilcox, 1998). For heavy-tailed symmetric distributions, Yuen (1974) found that a statistic based on trimmed means and Winsorized variances could adequately control the rate of Type I error and resulted in greater power than a statistic based on the usual mean and variance.

2.6 Scale Estimators

The value of a breakdown point is a main factor to be considered when looking for a scale estimator (Wilcox, 2005a). Rousseeuw and Croux (1993) have introduced several scale estimators with highest breakdown point, such as MAD_n, T_n and LMS_n. Due to their good performances in Huber (1981), Rousseeuw and Croux (1993) and Syed Yahaya *et al*. (2004a), these scale estimators are chosen for this study. All these scale estimators have 0.5 breakdown value and also exhibit bounded influence functions. These estimators are also chosen because of their simplicity and computational ease.

2.6.1 MAD_n

MAD_n is the median absolute deviation about the median. It demonstrates the best possible breakdown value of 50%, twice as much as the interquartile range and its influence function is bounded with the sharpest possible bound among all scale estimators (Rousseeuw & Croux, 1993).

This robust scale estimator is given by

$$MAD_n = b\,\mathrm{med}_i \left| x_i - med_j x_j \right|$$

where the constant b = 1.4826 is needed to make the estimator consistent for the parameter of interest, $x_i = x_1, x_2, ..., x_n$ and $i > j$

However, there are drawbacks in this scale estimator. The efficiency of MAD_n is very low with only 37% at Gaussian distribution. Rousseeuw and Croux (1993) have carried out a simulation on 10,000 batches of Gaussian observations to verify the efficiency gain at finite samples. They compared the variance of the standard deviation with the variance of *MAD_n* based on the finite samples. MAD_n also takes a symmetric view on dispersion and does not seem to be a natural approach for problems with asymmetric distributions.

2.6.2 T_n

Suitable for asymmetric distribution, Rousseeuw and Croux (1993) proposed T_n, a scale known for its highest breakdown point like MAD_n. However, this estimator has more plus points compared to MAD_n. It has 52% efficiency, making it more efficient than MAD_n. It also has a continuous and bounded influence function. Furthermore, the calculation of T_n is much easier than the other scale estimators.

Given as

$$T_n = 1.3800 \ \frac{1}{h} \sum_{k=1}^{h} \{\underset{j \neq i}{med} \mid x_i - x_j \mid\}_{(k).} \qquad \text{where } h = \left[\frac{n}{2}\right] + 1$$

T_n has a simple and explicit formula that guarantees uniqueness. This estimator also has 50% breakdown point.

2.6.3 LMS_n

LMS_n is also a scale estimator with a 50% breakdown point which is based on the length of the shortest half sample as shown below:

$$LMS_n = c' \min_i \left| x_{(i+h-1)} - x_{(i)} \right|$$

given $x_{(1)} \le x_{(2)} \le ... \le x_{(n)}$ are the ordered data and $h = \left[\frac{n}{2}\right] + 1$. The default value of c' is 0.7413 which achieves consistency at Gaussian distributions. *LMS_n* has an influence function as same as *MAD* (Rousseeuw & Leroy, 1987) and its efficiency equals that of the *MAD* as well (Grubel, 1988).

2.7 Statistical Methods

This study focuses on the T_1 statistic (Babu *et al.*, 1999) with several trimming criteria using robust scale estimators MAD_n, T_n and LMS_n. The Type I error rates and power of these tests under conditions of normality and nonnormality are examined. They were also compared to determine the best procedure and whether they were significantly better than the original T_1 statistic, with 15% symmetric trimming.

2.7.1 T_1 Statistic

When the distributions are symmetric, Babu *et al*. (1999) recommended the use of T_1 statistic to compare differences between distributions. They used a refined version of calculating trimmed means proposed by Rocke *et al.* (1982).

To calculate the T_1 statistic, let $X_{(1)j} \le X_{(2)j} \le ... \le X_{(n_j)j}$ represent the ordered observations associated with the *j*th group.

We calculated the *g*-trimmed mean of group *j*, $\overline{X}_{tj}$, by using:

$$\overline{X}_{tj} = \frac{1}{n_j - g_{1j} - g_{2j}} \left[\sum_{i=g_{1j}+1}^{n_j - g_{2j}} X_{(i)j} \right]$$

where

g_{1j} =number of observations $X_{(i)j}$ such that $\left(X_{(i)j} - \widehat{M}_j\right) <$ -2.24 (scale estimator),

g_{2j} =number of observations $X_{(i)j}$ such that $\left(X_{(i)j} - \widehat{M}_j\right) >$ 2.24 (scale estimator),

$\widehat{M}_j$ = median of group *j* and the scale estimator can be MAD_n, T_n and LMS_n.

n_j = group sample sizes

The value 2.24 was suggested by Wilcox and Keselman (2003) in place of the multiplier of the scale estimator in the above criteria. They denoted the multiplier with the letter *K*. They adjusted the *K* value so that efficiency is good under normality especially for small sample sizes. They found that, by using simulation with 10,000 replications, the efficiency of $\overline{X}_{tj}$ (the standard error of the sample mean divided by the standard error of $\overline{X}_{tj}$) is approximately 0.9 for $n_1 = n_2 = n_3 = n_4 = n_5 = 20$ with $K = 2.24$. $\overline{X}_{tj}$ was arrive at using MAD_n. We conducted a similar simulation study also on $\overline{X}_{tj}$ using robust scale estimators T_n and LMS_n, and found that the efficiencies are approximately 0.83 and 0.91, respectively. Hence, we kept the value of 2.24 in our selection criteria. Note that 2.24 is approximately equal to the square root of the 0.976 quantile of a chi-square distribution with one degree of freedom (Wilcox & Keselman, 2003). Indicating that, it is also suitable for skewed distribution.

The squared sample Winsorized standard error, $\hat{v}_{tj}{}^2$, is then defined as

$$\hat{v}_{tj}{}^2 = \frac{1}{(n_j - g_{1j} - g_{2j})(n_j - g_{1j} - g_{2j} - 1)} \times$$

$$\left[\sum_{i=g_{1j}+1}^{n_j - g_{2j}} \left(X_{(i)j} - \overline{X}_{tj} \right)^2 + g_{1j}(X_{(g_{1j}+1)j} - \overline{X}_{tj})^2 + g_{2j}(X_{(n_j - g_{2j})j} - \overline{X}_{tj})^2 \right].$$

Note that we used trimmed means in the $\widehat{v}_{tj}^2$ formula instead of Winsorized means.

Then the T_1 statistic is given by

$$T_1 = \sum_{1 \le j \le j' \le J} | t_{jj'} |,$$

where

$$t_{jj'} = \frac{\left(\overline{X}_{tj} - \overline{X}_{tj'} \right)}{\sqrt{\hat{v}_{tj} + \hat{v}_{tj'}}}$$

T$_1$ is the sum of all possible differences of sample trimmed means from the *J* distributions divided by their respective sample Winsorized standard errors. Therefore, if there are *J* distributions then the number of $t_{jj'}$s is equal to *J*(*J*-1)/2. Note that we used trimmed means in the Winsorized standard errors formula instead of Winsorized means.

2.8 Bootstrapping

The bootstrap is a Monte Carlo method that can be used to estimate the standard error of any estimator $\widehat{\theta}$ and was introduced by Efron (1979). The advantage of bootstrapping is its simplicity. This method is straightforward to apply to derive estimates of standard errors and confidence intervals for complex estimators of complex parameters of the distribution, such as percentile points, proportions, odds ratio, and correlation coefficients. Staudte and Sheather (1990) in their study stated that bootstrap is used to indicate that the observed data are used not only to obtain an estimate of the parameter but also to generate new samples. Bootstrap can routinely answer questions far too complicated for traditional statistical analysis. They work the same way (without formulae) for many different statistics in many different settings. In addition, bootstrapping can help in increasing accuracy of the test statistic.

When the sampling distribution of the estimator of interest is unknown, a pseudo sampling distribution of the estimator can be estimated using bootstrap. With the establishment of the pseudo sampling distribution, we can now assess variability of an estimator, bias of an estimator and significance of a test involving the estimator (Efron, 1979).

Bootstrap method is known to yield a better approximation than the one based on the normal approximation theory (Babu & Padmanabhan, 1996; Babu *et al.*, 1999). Othman *et al.* (2003) listed out two practical advantages of using bootstrap methods as detailed below;

i) Theory and empirical findings indicate that they can result in better Type I error control than non-bootstrap methods.

ii) Certain variations of the bootstrap method do not require the knowledge of the sampling distribution of the test statistic thereby not requiring explicit expressions for standard errors of estimators. This makes hypothesis testing quite flexible.

Westfall and Young (1993) suggested that Type I error control could be improved by combining bootstrap methods with methods based on trimmed means. The bootstrap seems preferable for general use if the goal is to avoid Type I error probability greater than the nominal level (Wilcox, 1998). The strategy behind the bootstrap is to use the shifted empirical distributions to estimate an appropriate critical value (Othman *et al*., 2003). Keseleman *et al.* (2003) stated that, further improvement in Type I error control is often possible by obtaining critical values for test statistic through bootstrap methods.

The bootstrap procedures on T_1 statistic is discussed in depth in Chapter 3.

CHAPTER 3

METHODOLOGY

3.1 Introduction

Our main focus in this study is to use robust scale estimators such as MAD_n, T_n and LMS_n as trimming criteria to trim data empirically. The mean of this trimmed data was calculated. A robust test statistic, namely T_1 were proposed to compare the differences between the groups regardless of assumptions. These statistics use group trimmed means as the central tendency measures. Unlike trimmed means that were based on trimmed observations of predetermined percentage, these proposed methods trimmed flexibly in order to avoid unnecessary trimming.

This study has been designed so as to encompass all conditions highlighting the strengths and weaknesses of testing the central tendency measures in achieving the objective of controlling Type I error and also the power of the tests. Under a completely randomized design, a few variables were manipulated to generate various conditions for testing robustness. The variables are the number of groups, balanced and unbalanced sample sizes, variance heterogeneity, nature of pairing of group variances and group sample sizes, and types of distributions. In addition to these variables, the study on power of tests also includes the setting of the central tendency parameters. This setting is dependent upon the effect size of the central tendency measures.

3.2 Procedures Employed

This study modified robust statistic, T_1 using automatic trimming strategy. These automatic trimming strategy involved robust scale estimators, MAD_n, T_n and LMS_n and predetermine asymmetric trimmed mean. We also include 15% symmetric trimming on T_1 as a benchmark. The 15% symmetric trimming in T_1 (Babu *et al.*, 1999) statistic produced good Type I error rates. Figure 3.1 shows the combinations of the statistical test with their corresponding scale estimators.

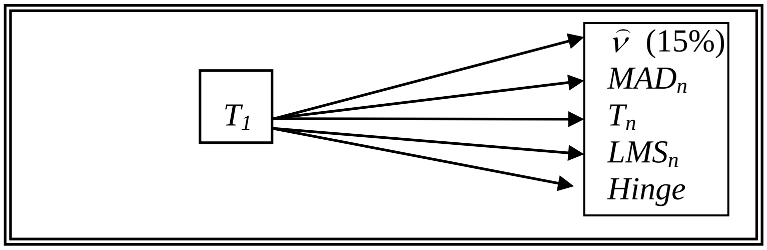

Figure 3.1: Statistical test with the corresponding scale estimators

These procedures are compared in terms of Type I error and power of the test statistics. Listed below are the four procedures:

i. T_1 with $\hat{\nu}$

ii. T_1 with MAD_n

iii. T_1 with T_n

iv. T_1 with LMS_n

v. T_1 with predetermine asymmetric trimming (*HQ1*)

vi. *t-test*

vii. *ANOVA*

The first (T_1 with $\hat{\nu}$) procedure is the benchmark procedures for the T_1 statistic. It is included in this study for comparison purposes.

Originally T_1 statistic used symmetric trimmed mean as their central tendency measure. This statistic works well under symmetric distribution. T_1 can handle variance heterogeneity and difference sample sizes. We still believe in trimmed mean. We believe that by trimming automatically, this statistic can handle heterogeneous, unequal sample sizes and skewed distribution. An additional feature of these new procedures includes the handling of the negative pairing of variances and sample sizes.

Rousseeuw and Croux (1993) suggested MAD_n, T_n and LMS_n as good robust scale estimators. Past studies by Wilcox and Keselman (2003) and Syed Yahaya (2005) have shown to provide good automatic trimming criterion for *MOM* based test statistic. Syed Yahaya (2005) has also shown similar results with T_n. Based on these; we can readily used MAD_n and T_n as automatic trimming criteria for trimmed mean. Since LMS_n is of the same class of robust scale estimator as MAD_n and T_n, it is an obvious choice as trimming criterion for automatic trimmed mean.

3.2.1 T_1 with $\hat{\nu}$

Let $X_{(1)j} \leq X_{(2)j} \leq \ldots \leq X_{(n_j)j}$ represent the ordered observations associated with the j^{th} group.

In order to calculate the 100g% sample trimmed mean, define

$$X_{Lj} = (1 - r)X_{(\mathrm{k}+1)j} + rX_{(\mathrm{k})j} \qquad [3.1]$$

and

$$X_{Uj} = (1 - r)\, X_{(n_j - k)j} + r\, X_{(n_j - k + 1)j} \qquad [3.2]$$

where

g represents the proportion of observations that are to be trimmed in each tail of the distribution.

$k = \lfloor gn_j \rfloor + 1$ where $\lfloor gn_j \rfloor$ is the largest integer $\leq gn_j$ and $r = k - gn_j$.

The *j*th group trimmed mean is given by

$$\overline{X}_{tj} = \frac{1}{(1-2g)n_j}\left[\sum_{i=k+1}^{n_j-k} X_{(i)j} + r(X_{(k)j} + X_{(n_j-k+1)j})\right] \qquad [3.3]$$

Its corresponding sample Winsorized mean is given by

$$\overline{X}_{wj} = \frac{1}{n_j}\left[\sum_{i=k+1}^{n_j-k} X_{(i)j} + k(X_{Lj} + X_{Uj})\right] \qquad [3.4]$$

The squared sample Winsorized standard error is as follows:

$$\hat{v}_{tj} = \frac{1}{(1-2g)n_j(n_j - 2n_jg - 1)} \times$$

$$\left[\sum_{i=k+1}^{n_j-k}\left(X_{(i)j} - \overline{X}_{wj}\right)^2 + k\left(\left(X_{Lj} - \overline{X}_{wj}\right)^2 + \left(X_{Uj} - \overline{X}_{wj}\right)^2\right)\right] \qquad [3.5]$$

Then the T_1 statistic is given by

$$T_1 = \sum_{1 \leq j \leq j' \leq J} | t_{jj'} |, \qquad [3.6]$$

where

$$t_{jj'} = \frac{\left(\overline{X}_{tj} - \overline{X}_{tj'}\right)}{\sqrt{\widehat{v}_{tj} + \widehat{v}_{tj'}}} \qquad [3.7]$$

3.2.2 T_1 with MAD_n

Let $X_{(1)j} \leq X_{(2)j} \leq \ldots \leq X_{(n_j)j}$ represent the ordered observations associated with the j^{th} group.

We calculated the g-trimmed mean of group j by using:

$$\overline{X}_{tj} = \frac{1}{n_j - g_{1j} - g_{2j}} \left[\sum_{i=g_{1j}+1}^{n_j - g_{2j}} X_{(i)j} \right] \qquad [3.8]$$

where

g_{1j} =number of observations $X_{(i)j}$ such that $\left(X_{(i)j} - \widehat{M}_j \right) <$ -2.24 (MAD_n)$_j$,

g_{2j} =number of observations $X_{(i)j}$ such that $\left(X_{(i)j} - \widehat{M}_j \right) >$ 2.24 (MAD_n)$_j$,

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

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

n_j = group sample sizes

The squared sample Winsorized standard error is then defined as

$$\hat{v}_{tj}^{\,2} = \frac{1}{(n_j - g_{1j} - g_{2j})(n_j - g_{1j} - g_{2j} - 1)} \times$$

$$\left[\sum_{i=g_{1j}+1}^{n_j - g_{2j}} \left(X_{(i)j} - \overline{X}_{tj} \right)^2 + g_{1j} (X_{(g_{1j}+1)j} - \overline{X}_{tj})^2 + g_{2j} (X_{(n_j - g_{2j})j} - \overline{X}_{tj})^2 \right] \qquad [3.9]$$

Then the T_1 statistic is given by

$$T_1 = \sum_{1 \leq j \leq j' \leq J} | t_{jj'} |, \qquad [3.10]$$

where

$$t_{jj'} = \frac{\left(\overline{X}_{tj} - \overline{X}_{tj'}\right)}{\sqrt{\widehat{\nu}_{tj} + \widehat{\nu}_{tj'}}}$$

T_1 is the sum of all possible differences of sample trimmed means from the J distributions divided by their respective sample Winsorized standard errors. Therefore, if there are J distributions then the number of $t_{jj'}$s is equal to J (J-1)/2. Note that trimmed means were used in the Winsorized standard errors formula instead of Winsorized means.

3.2.3 T_1 with T_n

We calculated the g-trimmed mean of group j by using:

$$\overline{X}_{tj} = \frac{1}{n_j - g_{1j} - g_{2j}} \left[\sum_{i=g_{1j}+1}^{n_j - g_{2j}} X_{(i)j} \right] \qquad [3.11]$$

where

g_{1j} =number of observations $X_{(i)j}$ such that $\left(X_{(i)j} - \widehat{M}_j\right) <$ -2.24 $(T_n)_j$,

g_{2j} =number of observations $X_{(i)j}$ such that $\left(X_{(i)j} - \widehat{M}_j\right) >$ 2.24 $(T_n)_j$,

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

$(T_n)_j$ = robust scale estimator of group j

n_j = group sample sizes

Then proceed with the calculation of $\overline{X}_{tj}$ and finally compute the T_1 statistic (equation [3.10]).

3.2.4 *T*₁ with *LMS*ₙ

We calculated the *g*-trimmed mean of group *j* by using:

$$\overline{X}_{tj} = \frac{1}{n_j - g_{1j} - g_{2j}} \left[\sum_{i=g_{1j}+1}^{n_j - g_{2j}} X_{(i)j} \right] \qquad [3.12]$$

where

g_{1j} =number of observations $X_{(i)j}$ such that $\left(X_{(i)j} - \widehat{M}_j\right) <$ -2.24 $(LMS_n)_j$,

g_{2j} =number of observations $X_{(i)j}$ such that $\left(X_{(i)j} - \widehat{M}_j\right) >$ 2.24 $(LMS_n)_j$,

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

$(LMS_n)_j$ = robust scale estimator of group *j*

n_j = group sample sizes

After the *g*-trimmed means were calculated, the compute $\overline{X}_{tj}$ followed by the corresponding T_1 statistic (equation [3.10]).

3.2.5 *Predetermined asymmetric trimmed mean (HQ1)*

The hypothesis to be tested in this report is

$$H_0: m_1 = m_2 = \ldots = m_j.$$

where m_j is adaptive trimmed mean for *j*th group and it is calculated as

$$m(\gamma_l, \gamma_u) = \frac{1}{h} \sum_{i=g_1+1}^{n_j - g_2} Y_i$$

where $g_1 = [n_j\gamma_l]$, $g_2 = [n_j\gamma_u]$, $h = n_j - g_1 - g_2$, γ_l = lower trimming percentage, γ_u = upper trimming percentage and n_j is the sample size. The percentage of lower and upper trimming identified using hinge estimator HQ_1 (Reed & Stark, 1996). However the total percentage of trimming is predetermined just like the usual trimmed mean.

To define the lower and upper trimming percentage, let consider an ordered sample J, L_α is the mean of the smallest $[\alpha n]$ observations, where $[\alpha n]$ denotes αn rounded down to the nearest integer, while U_α is the mean of the largest $[\alpha n]$ observations. As for example, let $\alpha = 0.05$, therefore $L_{0.05}$ is the mean of the smallest $0.05n$ observations. The measurement of Q_1 is defined as

$$Q_1 = \frac{U_{0.2} - L_{0.2}}{U_{0.5} - L_{0.5}}$$

Q_1 classifies whether a symmetric distribution has light (for $Q_1<2$), medium (for $2.6<Q_1\leq 3.2$) or heavy (for $Q_1>3.2$) tail. It is a location free statistic and uncorrelated with other location statistics. Reed and Stark (1996) defined a general scheme of their approach based on the former definitions of tail length as follows:

i. Set the total amount of trimming, γ, from the sample.

ii. Determine the proportion to be trimmed from the lower end of the sample (γl) by the proportion

$$\gamma_l = \gamma\left[\frac{UW_x}{UW_x + LW_x}\right]$$

where UW_x and LW_x are respectively the portion of the numerator and denominator of the previously defined statistic (Q_1). The notation for UW_x and LW_x are as follows:

$UW_{Q_1} = U_{0.2} - L_{0.2}$ and $LW_{Q_1} = U_{0.5} - L_{0.5}$

Subsequently, the calculation for HQ_1 is

$$HQ_1 = \frac{UW_{Q_1}}{UW_{Q_1} + LW_{Q_1}}$$

iii. The upper trimming percentage is defined as:

$$\gamma_u = \gamma - \gamma_l$$

3.3 Variables Manipulated

Several variables were manipulated in this study to create conditions which are known to emphasize the strengths and weaknesses of the proposed tests. This manipulation helps to identify the robustness of the tests in handling problems of nonnormality and heterogeneity.

3.3.1 Number of Groups

This study focused on a completely randomized design containing two and four groups (J = 2 and J = 4). Investigations on these two designs (J = 2 and J = 4) were chosen since previous work related to this study such as by Yuen (1974), Lix and Keselman (1998) and Othman *et al*. (2004) had also utilized similar designs. Furthermore, design containing four groups (J = 4) had been widely used successively among earlier researchers for its convincing F test results.

3.3.2 Balanced and Unbalanced Sample Sizes

For the purpose of examining the effect of sample sizes on Type I error and power of the investigated procedures, balanced and unbalanced sample sizes were allocated to each of the cases (J = 2 and J = 4). According to Othman *et al*. (2004), total sample sizes of 70 and 90 for J = 2 and J = 4 respectively, produced Type I error rates close to nominal value of α = 0.05, and it can be inferred that total sample sizes of any value within the 70 and 90 range should produce reasonably good Type I error rates (Syed Yahaya, 2005). Hence, in this study, total sample sizes for J = 4 was set at 60 and 80 while for the case of two groups (J = 2), the total sample sizes were half the values allocated to J = 4. Table 3.1 exhibits the sample sizes and variances for each group respectively.

Table 3.1: Sample sizes and variances

Groups	*Group Sizes (2 Cases)*		*Group Variances*
$J = 2$	N = 30	n_j = 15, 15	1:1
		n_j = 15, 15	1:36
		n_j = 12, 18	1:1
		n_j = 12, 18	1:36
		n_j = 12, 18	36:1
	N = 40	n_j = 20, 20	1:1
		n_j = 20, 20	1:36
		n_j = 15, 25	1:1
		n_j = 15, 25	1:36
		n_j = 15, 25	36:1
$J = 4$	N = 60	n_j = 15, 15, 15, 15	1:1:1:1
		n_j = 15, 15, 15, 15	1:1:1:36
		n_j = 12, 14, 16,18	1:1:1:1
		n_j = 12, 14, 16, 18	1:1:1:36
		n_j = 12, 14, 16, 18	36:1:1:1
	N = 80	n_j = 20, 20, 20, 20	1:1:1:1
		n_j = 20, 20, 20, 20	1:1:1:36
		n_j = 10, 20, 20, 30	1:1:1:1
		n_j = 10, 20, 20, 30	1:1:1:36
		n_j = 10, 20, 20, 30	36:1:1:1

For the purpose of comparison, this study covered both the (i) balanced and (ii) unbalanced sample sizes. For balanced sample sizes, the number of sample for J = 2 was set to be equal to 15 (n_1 = 15, n_2 = 15) and 20 (n_1 = 20, n_2 = 20) for each group. These values were also applied to both cases of J = 4 such that n_1 = 15, n_2 = 15, n_3 = 15, n_4 = 15 and n_1 = 20, n_2 = 20, n_3 = 20, n_4 = 20.

For the unbalanced case, each of the groups was arbitrarily assigned different numbers of observations, namely (i) n_1 = 12, n_2 = 18 and (ii) n_1 = 15, n_2 = 25 for J = 2. As for the four groups (J = 4) case, the smallest sample for N = 60 was set at 12 with subsequent increment of 2 for each group such that n_1 = 12, n_2 = 14, n_3 = 16, n_4 = 18 while for N = 80, the increment was not consistent for all the groups with n_1 = 10, n_2 = 20, n_3 = 20, n_4 = 30. Syed Yahaya *et al*. (2004a) and Keselman, Kowalchuk and Lix (1998) used N = 80 in their study and found that this number of sample sizes generated reasonable controlled Type I errors. Findings from Othman *et al*. (2004) also indicated that by using total sample sizes of 70 and 90, respectively the Type I error rates are close to the significance level ($\alpha = 0.05$). Therefore, it can be concluded that any value within the 70 and 90 as a total sample size should generate good Type I error rates. For two sample analysis with N = 30, Guo and Luh (2000) suggested the usage of the invertible Hall's transformation trimmed t under heterogeneity and nonnormality to get good control of Type I error rates. Therefore, this study adopted the same number of total sample sizes in evaluating the Type I error and power rates in order to achieve robust Type I error values. The chosen total sample sizes of N = 30 and N = 40 for J = 2 are half of the total sample sizes used for J = 4.

3.3.3 Variance Heterogeneity

The degree of variance heterogeneity is one of the factors affecting the robustness of the tests. In addition to affecting the Type I error rate, variance heterogeneity also affects the statistical power of the analysis (Wilcox, Charlin & Thomson, 1986). When the population variances differ, the actual statistical power can be less than that desired (Luh & Olejnik, 1990). To test for the effect of variance heterogeneity on Type I error and power, various ratios of variances were used in the literature.

Wilcox *et al*. (1986) have reported that the approximate tests still can provide liberal hypothesis tests if the ratio of standard deviations is as large as 4 to 1. In addition, Fenstad (1983) had chosen the ratio of 1:4 and stated the variances equal to 4 are not extreme as the ratio between the two standard deviations is only 2. Even though the ratio is quite small, there is a possibility of heteroscedastic variances. A ratio of 8:1 can be categorized as a less extreme ratio (Keselman *et al*., 2007). Report on real data by Higazi and Dayton (1984) uncovered that the estimated value of variance could exceed the value of 8. Wilcox (1987) stated that the value as large as 16 is not relatively common. He also said that some violation of the homogeneity of variance assumption is tolerable in terms of maintaining the nominal Type I error probability. Keselman, Lix *et al*. (1998) also found the usage of ratios 24:1 and 29:1 in a one-way and completely randomized factorial designs. Furthermore, Wilcox (2003) mentioned in his paper that a data set with variance ratio as high as 17977:1 was used in a study conducted by earlier researchers.

In this study, we looked at the effect of variance heterogeneity towards Type I error and power of the test statistics by using variance with ratio 1:36 (1:1:1:36). Even though the selected ratio seemed large, based on the previous literature, higher ratio than 1:36 (1:1:1:36) had been used by other researchers in their study (Keselman *et al*., 2007). In addition, it will also provide researchers with information regarding how well the methods hold up under any degree of heterogeneity they are likely to obtain in their data, thus providing a very generalizable result (Keselman *et al*., 2007). Othman *et al*. (2004) and Lix and Keselman (1998) also used the same ratio 36:1:1:1 in their research work to represent the extreme condition.

For the purpose of comparison, all the tests were conducted based upon similar conditions of variance homogeneity and heterogeneity. From this comparison, the tests that work excellently under this condition can be determined.

3.3.4 Nature of Pairings

When heterogeneous variances are paired with unbalanced sample sizes, there exist two types of pairings namely positive and negative pairings. Positive pairing refers to the case in which the largest sample size is associated with the population having the largest variance and the smallest sample size is associated with the population having the smallest variance. While negative pairing refers to the case in which the smallest sample size is associated with the population having the largest variance, and the largest sample size is associated with the population having the smallest variance.

Keselman, Othman *et al*. (2004), Othman *et al*. (2004), Syed Yahaya *et al*. (2004a) and Keselman *et al*. (2007) stated in their paper that the nature of pairings influenced the rate of Type I error. Syed Yahaya *et al*. (2004a) reported that for normal distribution, the difference in Type I error for each method they investigated was obvious for different pairings but for skewed distributions, the pairings did not show much difference in Type I error. In addition, Othman *et al*. (2004) wrote that empirically, the positive and negative pairings typically produce conservative and liberal Type I error rates, respectively.

Therefore, to appraise the robustness of the procedures in relation to the nature of the pairings, each of the proposed procedures was analyzed under the two types of pairings.

3.3.5 Types of Distributions

Practically a set of data will be nonnormally distributed when they are skewed or have heavy-tailed. In this study, types of distributions that can represent these two aspects of shape simultaneously will be generated to investigate the effects of distributional shape on Type I error and power. For that purpose, the *g*- and *h*-distribution is deemed to be the most suitable distribution in handling the skewness and the tail of a distribution. Introduced by Hoaglin (1985), this distribution provides a convenient method for considering a very wide range of situations corresponding to both symmetric and asymmetric distributions (Wilcox *et al.*, 2000).

In their study on the effect of distributional shape on Type I error and power, Othman *et al*. (2004), Wilcox (2005b) and Keselman *et al*. (2007) used data generated from the *g*- and *h*- distribution. The parameter *g*- controls the amount and the direction of skewness, while parameter *h*- controls the kurtosis or the amount of elongation. To observe the effect of distributional shapes on Type I error and power with regard to the procedures proposed, this study focused on three different shapes of distribution representing different degrees of skewness. The three distributions in ascending order of degree of skewness are normal, skewed normal-tail, and skewed leptokurtic (extremely skewed) represented by $g = h = 0$, $g = 0.5$ and $h = 0$, and $g = 0.5$ and $h = 0.5$ respectively.

To give meaning to these values, it should be noted that, for the standard normal distribution, $g = 0.0$ and $h = 0.0$. The zero value for g and h, respectively, signifies that the distribution is symmetric (no skew) and the tails are normally distributed. The tails of the distribution become heavier as h increases and are further skewed as g increases. This means that the g-distributions are more skewed to the right for larger values of g and positive values of h produce positive elongation. The larger the value of h, the more the elongation will be.

The next distribution used in this study is a skewed distribution with normal-tailed where $g = 0.5$ and $h = 0.0$. The skewness and kurtosis values for this distribution are $\gamma_1 = 1.75$ and $\gamma_2 = 8.9$, respectively (Keselman, Othman *et al*., 2004 & Othman *et al*., 2004).

For a more skewed effect, Algina, Penfield and Keselman (2005) suggested the application of $g = 0.225$ with skewness of $\gamma_1 = 4.9$ and $h = 0.225$ and kurtosis of $\gamma_2 = 4673.8$. Hence, the third distribution used in this study is the $g = 0.5$ dan $h = 0.5$. The values for the skewness and kurtosis for this distribution are undefined. Based on 100,000 observations, Wilcox (2005a, 2005b) reported the computer generated values for the skewness and kurtosis of this distribution as $\hat{\gamma}_1 = 120.10$ and $\hat{\gamma}_2 = 18393.6$, respectively. In this study, the case of $g = h = 0.5$ is being considered as to see how each method performs under an extreme condition. The idea is that if a method performs reasonably well under extreme conditions, this provides some assurance that it will perform well under conditions likely to be encountered in practice.

Table 3.2: Some Properties of the *g*- and *h*- Distribution

g	*h*	γ_1	γ_2	$\hat{\gamma}_1$	$\hat{\gamma}_2$
0.0	0.0	0.00	3.00	0.00	3.00
0.5	0.0	1.75	8.90	1.81	9.70
0.5	0.5	-	-	120.10	18393.6

Source: Wilcox (2005a)

Listed in Table 3.2 are the skewness (γ_1) and kurtosis (γ_2) for the three distributions considered here. The estimated skewness ($\hat{\gamma}_1$) and kurtosis ($\hat{\gamma}_2$) of the *g* = 0.5 and *h* = 0.5 distribution are based on 100,000 observations generated from the distribution are also shown in the table.

3.4 Design Specification

The design specifications shown in Table 3.3 – Table 3.10 are made up of the combinations of balanced and unbalanced sample sizes with homogenous and heterogeneous variances. The purpose of manipulating all the aforementioned variables is to examine the strengths and weaknesses of the tests. Design specifications for *J* = 2 and *J* = 4 are shown in the following tables.

Table 3.3: Design specification for equal sample sizes and homogeneous variances

(*J* = 2)

N = 30				*N* = 40			
Group Sizes		*Group Variances*		*Group Sizes*		*Group Variances*	
Group 1	Group 2	Group 1	Group 2	Group 1	Group 2	Group 1	Group 2
15	15	1	1	20	20	1	1

Table 3.4: Design specification for equal sample sizes and heterogeneous variances

(*J* = 2)

N = 30				*N* = 40			
Group Sizes		*Group Variances*		*Group Sizes*		*Group Variances*	
Group 1	Group 2	Group 1	Group 2	Group 1	Group 2	Group 1	Group 2
15	15	1	36	20	20	1	36

Table 3.5: Design specification for unequal sample sizes and homogeneous variances

(*J* = 2)

	N = 30				*N* = 40			
Pairing	*Group Sizes*		*Group Variances*		*Group Sizes*		*Group Variances*	
	Group 1	Group 2	Group 1	Group 2	Group 1	Group 2	Group 1	Group 2
	12	18	1	1	15	25	1	1

Table 3.6: Design specification for unequal sample sizes and heterogeneous variances

(*J* = 2)

	N = 30				*N* = 40			
Pairing	*Group Sizes*		*Group Variances*		*Group Sizes*		*Group Variances*	
	Group 1	Group 2	Group 1	Group 2	Group 1	Group 2	Group 1	Group 2
Positive	12	18	1	36	15	25	1	36
Negative	12	18	36	1	15	25	36	1

Table 3.7: Design specification for equal sample sizes and homogeneous variances

(*J* = 4)

N = 60							
Group Sizes				Group Variances			
Group 1	Group 2	Group 3	Group 4	Group 1	Group 2	Group 3	Group 4
15	15	15	15	1	1	1	1
N = 80							
Group Sizes				Group Variances			
Group 1	Group 2	Group 3	Group 4	Group 1	Group 2	Group 3	Group 4
20	20	20	20	1	1	1	1

Table 3.8: Design specification for equal sample sizes and heterogeneous variances

(*J* = 4)

N = 60							
Group Sizes				Group Variances			
Group 1	Group 2	Group 3	Group 4	Group 1	Group 2	Group 3	Group 4
15	15	15	15	1	1	1	36
N = 80							
Group Sizes				Group Variances			
Group 1	Group 2	Group 3	Group 4	Group 1	Group 2	Group 3	Group 4
20	20	20	20	1	1	1	36

Table 3.9: Design specification for unequal sample sizes and homogeneous variances

(*J* = 4)

N = 60							
Group Sizes				Group Variances			
Group 1	Group 2	Group 3	Group 4	Group 1	Group 2	Group 3	Group 4
12	14	16	18	1	1	1	1
N = 80							
Group Sizes				Group Variances			
Group 1	Group 2	Group 3	Group 4	Group 1	Group 2	Group 3	Group 4
10	20	20	30	1	1	1	36

Table 3.10: Design specification for unequal sample sizes and heterogeneous variances (J = 4)

Pairing	N = 60							
	Group Sizes				Group Variances			
	Group 1	Group 2	Group 3	Group 4	Group 1	Group 2	Group 3	Group 4
Positive	12	14	16	18	1	1	1	36
Negative	12	14	16	18	36	1	1	1
	N = 80							
	Group 1	Group 2	Group 3	Group 4	Group 1	Group 2	Group 3	Group 4
Positive	10	20	20	30	1	1	1	36
Negative	10	20	20	30	36	1	1	1

3.5 Data Generation

For each selected design specification shown in Table 3.3 – Table 3.10, five thousand data sets were simulated using significance level, α = 0.05. There are various numbers of simulations being used by previous researchers; usually, one thousand simulations were used for the trial stage and when the sampling distribution is known or be estimated. Ten thousand simulations were used when sampling distribution is really intractable or difficult to derive analytically. However, the frequently used number of simulations is five thousand.

In their work, Lix and Keselman (1998) used five thousand data sets to obtain Type I error rates for one-way completely randomized designs in which the underlying distributions were nonnormal, variances were nonhomogeneous and groups sizes were unequal. Keselman, Othman *et al*. (2004) also used five thousand replications for the new and improved two-sample *t* test. The same number of simulations was used by Syed Yahaya (2005) when examining Type I error and power rates of S_1 statistic with robust scale estimators MAD_n, S_n dan T_n.

The following steps will explain the generation of the pseudo-random variates for the *g*- and *h*- distribution:

(i) Generate standard normal variates, Z_{ij}. This involved a basic usage of SAS generator RANNOR (SAS Institute, 1999) with mean equals to zero and standard deviation equals to one.

(ii) Transform the standard normal variates to *g*- and *h*- variates via equation

$$X_{ij} = \begin{cases} \dfrac{\exp\left(gZ_{ij}{}^{2}\right)-1}{g}\exp\left(hZ_{ij}{}^{2}/2\right), & g \neq 0 \\ Z_{ij}\exp\left(hZ_{ij}{}^{2}/2\right), & g = 0 \end{cases} \qquad [3.20]$$

The parameter *g* controls the amount of skewness, while parameter *h* controls the kurtosis.

When dealing with skewed distributions, the central tendency measure such as the trimmed mean has value unequal to zero. To make certain that the null hypothesis remains true, the observations X_{ij}, from each simulated distributions were standardized by subtracting the population central tendency parameter, ω from the observations such that,

$$Y_{ij} = X_{ij} - \omega \qquad [3.21]$$

The value of ω were determined by computing $\hat{\omega}$ with one million observations generated from the distribution under study (Othman *et al.*, 2004; Wilcox & Keselman, 2003). Therefore, when working with trimmed mean, the population trimmed mean should be subtracted from X_{ij} to ensure that the null hypothesis for equal population trimmed means remains true.

According to the 1,000,000 observations generated for robust scale estimators MAD_n, LMS_n and also for $\hat{\nu}$, the population trimmed mean corresponding to the scale estimators for each type of distribution are listed in Table 3.11. While for robust scale estimator *T_n*, the population trimmed mean was generated from 100,000 observations only as the computing time for the scale estimator T_n was quite long. Due to the time constraint and the limited power of the computer, we decided that the values of the population trimmed mean for T_n were based on 100,000 observations only. These values are recorded in Table 3.11.

Table 3.11: Population trimmed mean for *g*- and *h*- distributions

Distributions	Robust scale estimators			
	MAD_n	T_n	LMS_n	$\widehat{\nu}$
g = 0.0 and *h* = 0.0	0.0021	0.0040	0.0023	0.0022
g = 0.5 and *h* = 0.0	0.0327	0.0286	0.2647	0.0772
g = 0.5 and *h* = 0.5	-0.0273	-0.0280	0.8075	0.0900

The observations are then transformed in accordance to the variance condition by multiplying $\sqrt{\sigma^2}$ to the centralized observations before adding in the term ω_j . Note that ω_j is the dispersion of group trimmed means which represents the degree of departure from no effect (null hypothesis).

For each design investigated, 5000 data sets were simulated. All of the procedures are tested using the 5% level of significance $(\alpha = 0.05)$. According to Manly (1997), for a test at 5% level of significance, the minimum 1000 data sets are almost certain to yield the same results as would a full distribution, However, when using 5000 data sets, 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 (Manly, 1997).

After trying out the simulation with 5000, 8000 and 10,000 data sets, the results showed that even though higher data sets were used the Type I error rates did not change much. Therefore, based on these finding, the minimum value of 5000 datasets were used as the number of randomizations. The simulated data sets are used to compute the test statistic, T_1.

For each procedure of T_1, every data set will then be bootstrapped B = 599 times. The choice of B = 599 is due to the fact that the test will produce the same result (converge to the same value) even if higher values are used. The effectiveness of this value was confirmed in Wilcox *et al*. (1998), who reported reasonable good results when using B = 599.

3.6 Percentile Bootstrap

Due to the intractability of the T_1 distribution, percentile bootstrap method was used to conduct the hypothesis test on the T_1 procedure. Babu *et al*. (1999) obtained the Type I error values for the S_1 and T_1 statistics by means of the percentile bootstrap method. They also discovered that the percentile bootstrap method produced better approximation than the one based on the normal approximation theory, and furthermore, this method works well especially when the samples are of moderate size.

3.6.1 T_1 with Bootstrap Method

To obtain the *p*-value of the T_1 statistic by using the percentile bootstrap method, the steps are as follows;

(a) Calculate T_1 based on the available data.

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

(c) Each of the sample points in the bootstrapped groups must be centered at their respective estimated trimmed means so that the sample trimmed mean is zero, such that $C^*_{ij} = X^*_{ij} - \overline{X}_{tj}$, $i = 1, 2, \ldots, n_j$. The empirical distributions are shifted so that the null hypothesis of the equal trimmed means among the *J* distributions is true. The strategy behind the bootstrap is to use the shifted empirical distributions to estimate an appropriate critical value.

(d) Let T^*_1 be the value of T_1 test based on the C^*_{ij} values.

(e) Repeat Step (a) to Step (d) *B* times yielding $T^*_{(1)1}, T^*_{(1)2}, \ldots, T^*_{(1)B}$. *B* = 599 appears sufficient in most situations when $n_j \geq 12$ (Wilcox, 2005a).

(f) Calculate the *p*-value as number of $\frac{T^*_{1B} > T_1}{B}$.

The calculated *p*-values are the estimated rates of Type I error for the procedures investigated under the T_1 statistic. The option of using *B* is based on Hall (1986; pp. 1453) which noted that:

> *To make our point about coverage probability, recall that if we conduct B bootstrap simulations, the resulting statistic values divide the real line into B + 1 parts. Therefore, in principle, confidence interval whose critical points are based on B simulations have coverage probabilities close to nominal levels* $\frac{b}{B+1}$ *for b = 1,…., B.*

Hall (1986) also stated that it is advantageous to choose *B* such that the nominal level, α , is a multiple of $(B+1)^{-1}$. Efron and Tibshirani (1993) suggested that *B* should be at least 500 or 1000 in order to make the variability of the estimated percentile acceptably low. When *B* is set at 599 instead of 600, the results show that the liberal values become non-liberal (Wilcox *et al*., 1998). They have found that the liberal values of Welch test decreased from 0.076 to 0.074 and from 0.078 to 0.077. Wilcox and Keselman (2002) have tried *B* = 2000 and they suggested that this number offers no practical value. In this study, *B* is set to be 599 with the reason that 599 is the lowest value that can make α a multiple of $(B+1)^{-1}$based on suggestion by Efron & Tibshirani (1993). Furthermore, trials on various numbers of bootstraps from *B* = 599 to 999 with the increment of 100 found that the *p*-values for different number of bootstraps are consistent. Thus, to save the running time, this study chose for the smallest *B* in the range.

CHAPTER 4
RESULTS OF THE ANALYSIS

4.1 Introduction

This study centered on the proposed test statistic for testing the equality of central tendency measures, namely T_1. This statistic was modified with predetermined asymmetric trimming and automatic trimming using robust scale estimators suggested by Rousseeuw and Croux (1993), i.e., MAD_n, T_n, and LMS_n. The procedures were then compared in terms of Type I error for their robustness. Various conditions to test the strengths and weaknesses of each of the procedures were also considered in this study such as the shape of distributions, balanced and unbalanced sample sizes, equal and unequal group variances, and the nature of the pairings of groups sample sizes and group variances. The procedures were then tested under two cases: the two (J = 2) and four (J = 4) groups cases. For each case, two total sample sizes (N) were suggested. As for J = 2, the total sample sizes are N = 30 and N = 40, while for J = 4, N = 60 and N = 80. The results, which are in the form of Type I error values, are presented in tables.

First column of the tables are types of distributions with different levels of skewness. They are g = 0.0 and h = 0.0, g = 0.5 and h = 0.0, and g = 0.5 and h = 0.5 representing zero, skewed with normal tail and extreme, respectively. For the unbalanced cases, the second column represents the nature of pairings (positive and negative) of group sample sizes and group variances. The other columns show the Type I error values obtained for each investigated procedure. These procedures are represented by their trimming strategies using robust scale estimators, namely MAD_n, T_n, LMS_n, $HQ1$ and 15% trimming, $\widehat{\nu}$. In this section, the T_1 procedures will be denoted by its estimator (e.g. T_1 with MAD_n is denoted as MAD_n). The rows represent the average values of each procedure corresponding to each distributional shape. The grand average values are displayed in the last row of every table. These grand average values were obtained by averaging all of the Type I error values generated by each procedure. Each value represents the overall performance of each procedure and all of the Type I error values reported are based on non-directional tests.

4.2 Type I Error for $J = 2$

The results of the analysis on the Type I error rates for $J = 2$ using T_1 statistic is shown in Table 4.1 to Table 4.4. The empirical Type I error rates are displayed for balanced and unbalanced sample sizes. For each condition, the table is partitioned according to the different total sample sizes, i.e., $N = 30$ and $N = 40$. The values that satisfied the Bradley's criterion of robustness are highlighted in bold and the average values that satisfied the criterion are also underlined.

4.2.1 Equal sample sizes and homogeneous variances

The empirical Type I error rates for the condition of equal sample sizes and homogenous variances are presented in Table 4.1 and Table 4.2.

Based on Table 4.1, the Type I error rates for $\hat{\nu}$, *t*-test and Mann-Whitney fall within the Bradley's interval regardless of distributions. For the proposed procedures, HQ1 also shows robustness except when the distribution is extremely skewed (g = 0.5, h = 0.5). The Mann- Whitney procedure produced the best Type I error rates under this condition.

As the sample size increased (*N* = 40), the Type I error rates for all procedures improved. The *HQ1* procedure robust throughout the three distributions. In addition, T_n also generate Type I error within Bradley's interval under normal distribution.

Table 4.1: Type 1 error rates (Equal sample sizes and homogeneous variances, *N* = 30)

N = 30 (15, 15) and variances (1:1)							
Distribution	T_1 with scale estimator						
	MAD_n	T_n	LMS_n	$\hat{\nu}$	*HQ1*	*t*-test	Mann-Whitney
g = 0.0 and h = 0.0	0.0170	0.0192	0.0130	**0.0454**	**0.0444**	**0.0462**	**0.0450**
g = 0.5 and h = 0.0	0.0138	0.0192	0.0104	**0.0420**	**0.0352**	**0.0426**	**0.0420**
g = 0.5 and h = 0.5	0.0088	0.0100	0.0076	**0.0372**	0.0172	**0.0276**	**0.0420**
Average	0.0132	0.0161	0.0103	**0.0415**	**0.0323**	**0.0388**	**0.0431**

Table 4.2: Type 1 error rates (Equal sample sizes and homogeneous variances, *N* = 40)

N = 40 (20, 20) and variances (1:1)							
Distribution	T_1 with scale estimator						
	MAD_n	T_n	LMS_n	$\hat{\nu}$	*HQ1*	*t*-test	Mann-Whitney
g = 0.0 and h = 0.0	0.0244	**0.0262**	0.0194	**0.0522**	**0.0536**	**0.0528**	**0.0560**
g = 0.5 and h = 0.0	0.0224	0.0234	0.0172	**0.0490**	**0.0486**	**0.0474**	**0.0566**
g = 0.5 and h = 0.5	0.0138	0.0150	0.0102	**0.0442**	**0.0336**	**0.0288**	**0.0526**
Average	0.0202	0.0215	0.0156	**0.0485**	**0.0452**	**0.0430**	**0.0539**

4.2.2 Equal sample sizes and heterogeneous variances

For this case, the empirical Type I error rates for all the procedures are displayed in Table 4.3 and Table 4.4. The Type I error rates for $\hat{\nu}$, *HQ1* and *t*-test are robust regardless of distributions. The automatic trimming procedure T_n is also robust except under extreme distribution. The results show some improvement in the automatic trimming of T_n and MAD_n, as the sample size increased.

Table 4.3: Type 1 error rates (Equal sample sizes and heterogeneous variances *N* = 30)

N = 30 (15, 15) and variances = (1:36)							
Distribution	T_1 with scale estimator						
	MAD_n	T_n	LMS_n	$\hat{\nu}$	*HQ1*	*t*-test	Mann-Whitney
g = 0.0 and h = 0.0	0.0248	**0.0272**	0.0238	**0.0468**	**0.0624**	**0.0596**	0.0766
g = 0.5 and h = 0.0	**0.0332**	**0.0326**	0.0808	**0.0518**	**0.0432**	**0.0746**	0.0776
g = 0.5 and h = 0.5	0.0164	0.0144	0.2666	**0.0482**	**0.0416**	**0.0374**	**0.0664**
Average	0.0248	0.0247	0.1230	**0.0489**	**0.0491**	**0.0572**	**0.0735**

Table 4.4: Type 1 error rates (Equal sample sizes and heterogeneous variances *N* = 40)

N = 40 (20, 20) and variances = (1:36)							
Distribution	T_1 with scale estimator						
	MAD_n	T_n	LMS_n	$\hat{\nu}$	*HQ1*	*t*-test	Mann-Whitney
g = 0.0 and h = 0.0	**0.0298**	**0.0320**	**0.0278**	**0.0538**	**0.0592**	**0.0618**	0.0938
g = 0.5 and h = 0.0	**0.0414**	**0.0448**	0.1132	**0.0568**	**0.0602**	0.0882	0.0820
g = 0.5 and h = 0.5	0.0192	0.0214	0.4228	**0.0516**	**0.0676**	**0.0430**	0.0758
Average	**0.0301**	**0.0327**	0.1879	**0.0541**	**0.0623**	**0.0643**	0.0853

4.2.3 Unequal sample sizes and homogeneous variances

The empirical Type I error rates for unequal sample sizes and homogeneous variances are shown in Table 4.5 and Table 4.6. The group sample sizes for N = 30 and N = 40 were set as n_1 = 12, n_2 = 18 and n_1 = 15, n_2 = 25 respectively.

Table 4.5: Type 1 error rates (Unequal sample sizes and homogeneous variances, N = 30)

N = 30 (12, 18) and variances = (1:1)							
Distribution	T_1 with scale estimator						
	MAD_n	T_n	LMS_n	$\widehat{\nu}$	*HQ1*	*t*-test	Mann-Whitney
g = 0.0 and h = 0.0	0.0200	0.0222	0.0170	**0.0454**	**0.0482**	**0.0494**	**0.0502**
g = 0.5 and h = 0.0	0.0190	0.0220	0.0148	**0.0430**	**0.0344**	**0.0476**	**0.0472**
g = 0.5 and h = 0.5	0.0100	0.0110	0.0076	**0.0368**	0.0162	**0.0332**	**0.0472**
Average	0.0163	0.0184	0.0131	**0.0417**	**0.0329**	**0.0434**	**0.0482**

Table 4.6: Type 1 error rates (Unequal sample sizes and homogeneous variances, N = 40)

N = 40 (15, 25) and variances = (1:1)							
Distribution	T_1 with scale estimator						
	MAD_n	T_n	LMS_n	$\widehat{\nu}$	*HQ1*	t-test	Mann-Whitney
g = 0.0 and h = 0.0	**0.0250**	**0.0258**	0.0216	**0.0512**	**0.0496**	**0.0490**	**0.0510**
g = 0.5 and h = 0.0	0.0238	**0.0272**	0.0218	**0.0492**	**0.0408**	**0.0468**	**0.0502**
g = 0.5 and h = 0.5	0.0170	0.0172	0.0122	**0.0450**	0.0218	**0.0324**	**0.0520**
Average	0.0219	0.0234	0.0185	**0.0485**	**0.0374**	**0.0427**	**0.0511**

For both tables, every entry in the $\widehat{\nu}$, *t*-test and Mann-Whitney is highlighted in bold, which reflects that the procedure has good control of Type I error rates across the three types of distributions for both sample sizes. The proposed *HQ1* is in control of

Type I error rates under normal and mildly skewed distributions. However, none of the automatic trimming procedures are robust for small sample size, but as the sample size increased, T_n and MAD_n under normal distribution become robust, while T_n maintains its robustness even under mildly skewed distributions.

4.2.4 Unequal sample sizes and heterogeneous variances

The results of the investigation on unequal sample sizes and heterogeneous variances are presented in Table 4.7 and Table 4.8. For this case, as stated earlier, there is an additional column for the pairing category. Positive pairing refers to the case in which the largest sample size is associated with population having the largest variance and the smallest sample size is associated with the population having the smallest variance. While negative pairing refers to the case in which the smallest sample size is associated with the population having the largest variance, and the largest sample size is associated with the population having the smallest variance.

As shown in Table 4.7 and Table 4.8, the $\hat{v}$ and *HQ1* are robust across the three types of distributions for both total sample sizes. However, Mann-Whitney only produced robust Type I error rates for positive pairing regardless of type of distribution.

Table 4.7: Type 1 error rates (Unequal sample sizes and heterogeneous variances, N = 30)

N = 30 (12, 18) and variances (1:36) and (36:1)								
Distribution	Pairing	T_1 with scale estimator						
		MAD_n	T_n	LMS_n	$\hat{\nu}$	*HQ1*	*t*-test	Mann-Whitney
g = 0.0 and h = 0.0	Positive	0.0246	**0.0264**	0.0226	**0.0504**	**0.0546**	0.0242	**0.0532**
	Negative	**0.0256**	0.0220	0.0236	**0.0462**	**0.0572**	0.1168	0.1202
Average		**0.0251**	0.0242	0.0231	**0.0483**	**0.0559**	**0.0705**	0.0867
g = 0.5 and h = 0.0	Positive	**0.0350**	**0.0340**	0.0934	**0.0500**	**0.0524**	**0.0276**	**0.0536**
	Negative	**0.0376**	**0.0334**	0.0724	**0.0506**	**0.0392**	0.1394	0.1092
Average		**0.0363**	**0.0337**	0.0829	**0.0503**	**0.0458**	0.0835	0.0814
g = 0.5 and h = 0.5	Positive	0.0164	0.0158	0.3530	**0.0510**	**0.0384**	0.0124	**0.0458**
	Negative	0.0168	0.0132	0.1996	**0.0450**	**0.0374**	0.0892	0.0968
Average		0.0166	0.0145	0.2763	**0.0480**	**0.0379**	**0.0508**	**0.0713**
Grand Average		**0.0260**	0.0241	0.1294	**0.0489**	**0.0465**	**0.0683**	0.0798

Table 4.8: Type 1 error rates (Unequal sample sizes and heterogeneous variances, N = 40)

N = 40 (15, 25) and variances (1:36) and (36:1)								
Distribution	Pairing	T_1 with scale estimator						
		MAD_n	T_n	LMS_n	$\hat{\nu}$	*HQ1*	*t*-test	Mann-Whitney
g = 0.0 and h = 0.0	Positive	**0.0326**	**0.0344**	**0.0274**	**0.0548**	**0.0620**	**0.0270**	**0.0508**
	Negative	0.0242	**0.0256**	0.0244	**0.0466**	**0.0594**	0.1290	0.1244
Average		**0.0284**	**0.0300**	**0.0259**	**0.0507**	**0.0607**	0.0780	0.0876
g = 0.5 and h = 0.0	Positive	**0.0380**	**0.0416**	0.1302	**0.0540**	**0.0604**	**0.0340**	**0.0440**
	Negative	**0.0352**	**0.0384**	0.0858	**0.0466**	**0.0400**	0.1540	0.1068
Average		**0.0366**	**0.0400**	0.1080	**0.0503**	**0.0502**	0.0940	0.0754
g = 0.5 and h = 0.5	Positive	0.0186	0.0216	0.5230	**0.0552**	**0.0648**	0.0140	**0.0420**
	Negative	0.0172	0.0162	0.2624	**0.0470**	**0.0382**	0.1020	0.1080
Average		0.0179	0.0189	0.3927	**0.0511**	**0.0515**	**0.0580**	**0.0750**
Grand Average		**0.0276**	**0.0296**	0.1755	**0.0507**	**0.0541**	0.0767	0.0793

For T_1 with automatic trimming strategies, the results are not consistent. Under normal distribution with N = 30, only MAD_n shows robustness based on the average value. As the total sample size increased, the performance of T_n and LMS_n also improved where all the automatic trimming strategies are robust.

When the distribution is skewed with normal tail (g = 0.5 and h = 0.0), the Type I error rates for T_n and MAD_n are within the Bradley's interval for both pairings and both total sample sizes, but this is not the case for LMS_n. As the skewness increased, the Type I error rates for the automatic trimming worsen. None of the procedures can be considered robust under extremely skewed distribution. The *t*-test does not seem to perform well under most conditions.

4.3 Type 1 Error for *J* = 4

The previous section (4.2.1) discussed on the results of J = 2 case. Like in the case of J = 2, the conditions for the four groups case (J = 4) are the same except for changes in the total sample sizes to N = 60 and N = 80. The results of the analysis of the Type I error rates using T_1 statistic for J = 4 are shown in Table 4.9 to Table 4.16.

4.3.1 Equal sample sizes and homogeneous variances

For the condition of equal sample sizes and homogeneous variances, the results for all the procedures investigated are presented in Table 4.9 and Table 4.10. Like in the previous sections, the values that satisfied the Bradley's robust criterion were highlighted in bold and the average values that satisfied the criterion were also highlighted and underlined.

Based on all of the Type I error values in Table 4.9 and Table 4.10, T_1 with *HQ1*, T_1 with $\hat{\nu}$, ANOVA and Kruskal Wallis produced Type I error rates that satisfied Bradley's criterion of robustness. As sample sizes increased the average Type I error values for all the procedures improved (approaching 0.05). However, the automatic trimming strategies still produced conservative Type I error for T_1 statistic even with larger sample sizes.

Table 4.9: Type 1 error rates (Equal sample sizes and homogeneous variances, N = 60)

N = 60 (15, 15, 15, 15) and variances (1:1:1:1)							
Distribution	T_1 with scale estimator						
	MAD_n	T_n	LMS_n	$\hat{\nu}$	*HQ1*	ANOVA	Kruskal Wallis
g = 0.0 and h = 0.0	0.0062	0.0084	0.0028	**0.0404**	**0.0434**	**0.0498**	**0.0418**
g = 0.5 and h = 0.0	0.0050	0.0080	0.0022	**0.0366**	**0.0286**	**0.0490**	**0.0448**
g = 0.5 and h = 0.5	0.0030	0.0030	0.0016	**0.0294**	**0.0072**	**0.0274**	**0.0448**
Average	0.0047	0.0065	0.0022	**0.0355**	**0.0264**	**0.0421**	**0.0438**

Table 4.10: Type 1 error rates (Equal sample sizes and homogeneous variances, N = 80)

N = 80 (20, 20, 20, 20) and variances (1:1:1:1)							
Distribution	T_1 with scale estimator						
	MAD_n	T_n	LMS_n	$\hat{\nu}$	*HQ1*	*ANOVA*	Kruskal Wallis
g = 0.0 and h = 0.0	0.0110	0.0124	0.0096	**0.0452**	**0.0486**	**0.0518**	**0.0464**
g = 0.5 and h = 0.0	0.0114	0.0116	0.0068	**0.0402**	**0.0412**	**0.0550**	**0.0498**
g = 0.5 and h = 0.5	0.0050	0.0044	0.0028	**0.0318**	0.0208	**0.0290**	**0.0498**
Average	0.0091	0.0095	0.0064	**0.0391**	**0.0369**	**0.0453**	**0.0487**

4.3.2 Equal sample sizes and heterogeneous variances

Table 4.11 and Table 4.12 depicts the empirical Type I error rates for the condition of equal sample sizes and heterogeneous variances. The group variances were set at 1:1:1:36.

Table 4.11: Type 1 error rates (Equal sample sizes and heterogeneous variances, N = 60)

N = 60 (15, 15, 15, 15) and variances (1:1:1:36)							
Distribution	T_1 with scale estimator						
	MAD_n	T_n	LMS_n	$\widehat{\nu}$	*HQ1*	ANOVA	Kruskal Wallis
g = 0.0 and h = 0.0	0.0124	0.0156	0.0072	**0.0410**	**0.0526**	0.1046	**0.0690**
g = 0.5 and h = 0.0	0.0176	0.0216	**0.0388**	**0.0424**	**0.0386**	0.1380	**0.0696**
g = 0.5 and h = 0.5	0.0068	0.0068	0.1514	**0.0368**	0.0202	0.2270	**0.0642**
Average	0.0123	0.0147	**0.0658**	**0.0401**	**0.0371**	0.1565	**0.0676**

Table 4.12: Type 1 error rates (Equal sample sizes and heterogeneous variances, N = 80)

N = 80 (20, 20, 20, 20) and variances (1:1:1:36)							
Distribution	T_1 with scale estimator						
	MAD_n	T_n	LMS_n	$\widehat{\nu}$	*HQ1*	*ANOVA*	Kruskal Wallis
g = 0.0 and h = 0.0	0.0194	0.0214	0.0162	**0.0498**	**0.0548**	0.1096	**0.0720**
g = 0.5 and h = 0.0	**0.0262**	**0.0258**	**0.0672**	**0.0490**	**0.0532**	0.1270	**0.0734**
g = 0.5 and h = 0.5	0.0110	0.0102	0.3006	**0.0468**	**0.0544**	0.2346	**0.0684**
Average	0.0189	0.0191	0.1280	**0.0485**	**0.0541**	0.1571	**0.0713**

A glance through the columns shows that, the T_1 statistic with automatic trimming strategies using MAD_n and T_n produced conservative Type I error rates, while for the other trimming strategy, i.e. LMS_n, we observe an erratic pattern of the values. In contrast, ANOVA produced liberal values exceeding 0.1. Robust values could be observed under the columns of $\hat{\nu}$ and Kruskal Wallis for sample sizes N = 60. Even though *HQ1* produced average robust value, but this procedure failed to control its Type I error under extreme condition (g = 0.5 and h = 0.5). When the sample size was increased to 80, the rates under extreme condition improved tremendously. Consistent values were observed under *HQ1*. The larger sample size also improved the Type I error rates for $\hat{\nu}$. Even though the values for Kruskal Wallis procedure are still within the robust criteria, the effect of larger sample size also caused the Type I error values to inflate.

4.3.3 Unequal sample sizes and homogeneous variances

The performance of the procedures under unequal sample sizes and homogeneous variances are shown in Table 4.13 and Table 4.14.

Under this condition, the $\hat{\nu}$, ANOVA and Kruskal Wallis column produced Type I error values which are within the Bradley's interval for both total sample sizes. When the total sample size increased from N = 60 to N = 80, the Type I error values for $\hat{\nu}$, ANOVA and Kruskal Wallis are also improved, producing Type I error values which are nearer to the nominal level $(\alpha = 0.05)$. The Type I error value for *HQ1* are also robust for normal distribution and still robust when the distribution is skewed with normal tailed (g = 0.5 and h = 0.0). Unlike $\hat{\nu}$, ANOVA and Kruskal Wallis, the Type I error values for *HQ1* are decreased when the number of sample size are increased.

Table 4.13: Type 1 error rates (Unequal sample sizes and homogeneous variances, N = 60)

N = 60 (12, 14, 16, 18) and variances (1:1:1:1)							
Distribution	T_1 with scale estimator						
	MAD_n	T_n	LMS_n	$\hat{\nu}$	*HQ1*	*ANOVA*	Kruskal Wallis
g = 0.0 and h = 0.0	0.0108	0.0110	0.0050	**0.0452**	**0.0406**	**0.0484**	**0.0440**
g = 0.5 and h = 0.0	0.0060	0.0052	0.0030	**0.0370**	**0.0254**	**0.0444**	**0.0422**
g = 0.5 and h = 0.5	0.0028	0.0024	0.0020	**0.0272**	0.0058	**0.0276**	**0.0422**
Average	0.0065	0.0062	0.0033	**0.0365**	0.0239	**0.0401**	**0.0428**

Table 4.14: Type 1 error rates (Unequal sample sizes and homogeneous variances, N = 80)

N = 80 (10, 20, 20, 30) and variances (1:1:1:1)							
Distribution	T_1 with scale estimator						
	MAD_n	T_n	LMS_n	$\hat{\nu}$	*HQ1*	*ANOVA*	Kruskal Wallis
g = 0.0 and h = 0.0	0.0168	0.0150	0.0142	**0.0486**	**0.0498**	**0.0492**	**0.0420**
g = 0.5 and h = 0.0	0.0132	0.0160	0.0110	**0.0464**	**0.0372**	**0.0494**	**0.0506**
g = 0.5 and h = 0.5	0.0074	0.0068	0.0056	**0.0334**	0.0162	**0.0388**	**0.0506**
Average	0.0125	0.0126	0.0103	**0.0428**	**0.0344**	**0.0458**	**0.0477**

4.3.4 Unequal sample sizes and heterogeneous variances

The Type I error rates presented in Table 4.15 and Table 4.16 were obtained from the tests performed on the unequal sample sizes and heterogeneous group variances. For this case, the study also involved investigations on the positive and negative pairings of group variances and group sample sizes. The Type I error values for both pairings were averaged and recorded under "Average" for each distribution. Table 4.15 and Table 4.16 also include the "Grand Average" which represents the average Type I error values across the distributions.

As shown in the table, all the Type I error values of *T_1* statistic which used $\widehat{v}$ as trimming strategy are robust according to the Bradley's interval for both total sample sizes. The values ranging from 0.0376 to 0.0506 are quite close to the nominal level $(\alpha = 0.05)$. The Type I error values for *HQ1* are also robust for normal distribution and *g* = 0.5 and *h* = 0.0 distribution. In contrast, *T_1* with *MAD_n*, *T_n* and *LMS_n*, *ANOVA* and Kruskal Wallis produced Type I error that met Bradley's criterion for certain cases only. They are i) under skewed normal-tailed distribution and negative pairing for both sample sizes, Type I error rates that satisfied the Bradley's criterion were observed, ii) again, under skewed normal-tailed distribution and positive pairing for both sample sizes, *T_1* with *LMS_n* produced Type I error rates within Bradley's criterion, iii) for *ANOVA,* under normal distribution for both total sample sizes, iv) robust Type I error values only on positive pairing for both sample sizes for Kruskal Wallis.

Table 4.15: Type 1 error rates (Unequal sample sizes and heterogeneous variances, N = 60)

N = 60 (12, 14, 16, 18) and variances (1:1:1:36) and (36:1:1:1)								
Distribution	Pairing	T_1 with scale estimator						
		MAD_n	T_n	LMS_n	$\hat{\nu}$	*HQ1*	*ANOVA*	Kruskal Wallis
g = 0.0 and h = 0.0	Positive	0.0112	0.0116	0.0082	**0.0438**	**0.0398**	**0.0678**	**0.0584**
	Negative	0.0172	0.0146	0.0164	**0.0428**	**0.0550**	0.1596	0.0832
Average		0.0142	0.0131	0.0123	**0.0433**	**0.0474**	0.1137	**0.0708**
g = 0.5 and h = 0.0	Positive	0.0174	0.0164	**0.0434**	**0.0406**	**0.0394**	0.0978	**0.0596**
	Negative	**0.0260**	0.0248	**0.0492**	**0.0454**	**0.0342**	0.2002	0.0848
Average		0.0217	0.0206	**0.0463**	**0.0430**	**0.0368**	0.1490	**0.0722**
g = 0.5 and h = 0.5	Positive	0.0066	0.0062	0.2094	**0.0376**	0.0174	0.1920	**0.0520**
	Negative	0.0098	0.0086	0.1422	**0.0392**	0.0224	0.2628	0.0760
Average		0.0082	0.0074	0.1758	**0.0384**	0.0199	0.2274	**0.0640**
Grand Average		0.0147	0.0137	0.0781	**0.0416**	**0.0347**	0.1634	**0.0690**

Table 4.16: Type 1 error rates (Unequal sample sizes and heterogeneous variances, N = 80)

N = 80 (10, 20, 20, 30) and variances (1:1:1:36) and (36:1:1:1)								
Distribution	Pairing	T_1 with scale estimator						
		MAD_n	T_n	LMS_n	$\hat{\nu}$	*HQ1*	*ANOVA*	Kruskal Wallis
g = 0.0 and h = 0.0	Positive	0.0184	0.0174	0.0128	**0.0504**	**0.0616**	**0.0336**	**0.0418**
	Negative	0.0206	0.0208	0.0176	**0.0464**	**0.0476**	0.2840	0.1148
Average		0.0195	0.0191	0.0152	**0.0484**	**0.0546**	0.1588	0.0783
g = 0.5 and h = 0.0	Positive	0.0190	0.0218	**0.0702**	**0.0470**	**0.0394**	0.0754	**0.0478**
	Negative	**0.0252**	**0.0252**	**0.0476**	**0.0506**	**0.0354**	0.3282	0.1152
Average		0.0221	0.0235	**0.0589**	**0.0488**	**0.0374**	0.2018	0.0815
g = 0.5 and h = 0.5	Positive	0.0074	0.0096	0.3896	**0.0424**	**0.0288**	0.1486	**0.0446**
	Negative	0.0128	0.0096	0.1240	**0.0410**	**0.0300**	0.3544	0.0996
Average		0.0101	0.0096	0.2567	**0.0417**	**0.0294**	0.2515	**0.0481**
Grand Average		0.0172	0.0174	0.1103	**0.0463**	**0.0405**	0.2040	**0.0693**

CHAPTER 5
DISCUSSION AND CONCLUSIONS

5.1 Introduction

The goal of this study was to find alternative methods in testing the equality of location measures as most frequently used conventional methods such as the ANOVA *F*-test or the *t*-test are known to be sensitive to certain assumptions. In particular, non-normality and variance heterogeneity tend to disrupt the Type I error control and the power to detect differences between the central tendency measures. Other alternative methods namely the nonparametric procedures are known to be less powerful.

In this study, a test statistics for testing the equality of central tendency measures was proposed. The procedure is the T_1 statistic by itself, not in the adaptive manner as originally introduced by Babu *et al*. (1999). T_1 is a statistic suitable to be used when the distributions are symmetric. In its original state, T_1 statistics used trimmed mean as the central tendency measure. In this study, we proposed that the trimmed mean for T_1 is obtained via three types of trimming method; (i) automatic trimming strategy which the amount of trimming is determined by the characteristics of the sample data, (ii) predetermined asymmetric trimming based upon hinge estimators and (iii) fix amount of trimming. The automatic trimming strategy was based upon trimming criteria using robust scale estimators, MAD_n, T_n and LMS_n. These estimators were recommended by Rousseeuw and Croux (1993) due to their highest breakdown value (0.5) and bounded influence function. In addition, these estimators are simple and easy to compute. The fix amount of trimming, 15% symmetric trimming ($\hat{\nu}$) was used as a benchmark in this study. This study also included *t*-test and ANOVA and also nonparametric methods, such as Mann-Whitney and Kruskal Wallis.

Altogether, seven procedures were proposed. The modified procedures under the T_1 statistic are T_1 statistic with automatic trimming using robust scale estimators, MAD_n, T_n and LMS_n, the T_1 statistic with predetermined asymmetric trimming. The T_1 with $\hat{\nu}$, t-test or ANOVA and Mann-Whitney or Kruskal Wallis procedures were considered in this study for comparison purposes.

Each of the proposed procedures was then tested for the effect of non-normality and heteroscedasticity on the Type I error and power of test. To accomplish the tests, three types of distributional shapes representing different levels of skewness were used to test the non-normality effect and the unequal variances of 1:36 ratio were used for the heteroscedasticity effect. In addition other variables such as the number of groups (varying from two to four), and the nature of pairings of group sizes and group variances (positive and negative pairing) were also included as these variables were also proven to have some effect on the rates of Type I error and power of test (Othman *et al*., 2004). For the purpose of comparison, Type I error rates were also collected for the balanced design consisting of equal sample sizes and equal group variances.

Data from the *g*- and *h*- distributions representing different level of skewness and kurtosis were used to test the nonnormality effect. The normal distribution is represented by the distribution of $g = 0.0$ and $h = 0.0$, the skewed normal-tailed is represented by the $g = 0.5$ and $h = 0.0$ distribution, while the $g = 0.5$ and $h = 0.5$ represents the skewed leptokurtic (extremely skewed) distribution. Each procedure was simulated 5000 times and then bootstrapped 599 times. Due to the intractability of the sampling distributions of the statistics, the bootstrap percentile method was used to test the hypothesis.

The rates of Type I error for each procedure were determined and compared. In searching for the best procedure/s, comparisons were made between the modified procedures versus the original procedures, and also within the modified procedures. The best procedure will produce the empirical Type I error rate nearest to the nominal value of 0.05. The robustness of the procedures was also examined. By adopting Bradley's (1978) liberal criterion of robustness, a procedure is considered robust if its empirical rate of Type I error, is within the interval 0.025 and 0.075 for a 0.05 significance level.

5.2 The new T_1 procedures

5.2.1 Case of two groups

Under ideal condition (equal sample sizes, homogeneous variance, and normal distribution), the conventional (*t*-test and Mann-Whitney), the original ($\hat{\nu}$), and the *HQ1* produced robust Type I error rates. T_n also showed robustness, but under larger sample size. The performance of the conventional, original and *HQ1* maintain even under skewed distributions, but none of the automatic trimming procedures achieved robustness.

When variances are heterogenous and sample sizes are equal, most of the procedures under normal distribution are robust especially for larger sample size. However, for both sample sizes, the Mann-Whitney procedure is not robust. Under skewed distribution, the proposed automatic trimming procedures become better except for LMS_n.

Under unequal sample sizes with homogeneous variances, the conventional and original procedures have good control of Type I error rates for both total sample sizes. The proposed asymmetric trimming, *HQ1* performs as good as the conventional and original procedures, but not when the distribution is extremely skewed. Even though fail to perform under smaller sample size, the automatic trimming procedures namely MAD_n and T_n improved when the sample size increased.

In the case of unequal sample sizes and heterogenous variances, the *HQ1* performs as good as the original procedure where all the Type I error are within the robust criterion. The Type I error rates for MAD_n and T_n are also within the robust criterion under normal and mildly skewed. In general under this case, most of the Type I error rates of the conventional procedures are not within the robust interval with values as large as 0.1540 for *t*-test and 0.1244 for Mann-Whitney.

5.2.2 Case of four groups

For conditions of equal sample sizes, homogeneous variance, and normal distribution, the conventional (ANOVA and Kruskal Wallis), the original ($\hat{\nu}$), and the *HQ1* produced robust Type I error rates. The performance of the conventional, original and *HQ1* maintain even under skewed distributions, but none of the automatic trimming procedures achieved robustness.

Under equal sample sizes and variances are heterogenous, the original, Kruskal Wallis and the *HQ1* procedures maintain their performances regardless of distributions. LMS_n also showed robustness for mild skewed distribution. MAD_n and T_n also produced robust Type I error under larger sample size.

When variances are homogeneous and sample sizes are unequal, the conventional and original procedures have good control of Type I error rates for both total sample sizes. The proposed asymmetric trimming, *HQ1* performs as good as the conventional and original procedures, but not when the distribution is extremely skewed.

In the case of unequal sample sizes and heterogenous variances, the *HQ1* performs as good as the original procedure where all the Type I error are within the robust criterion except for smaller sample sizes under extremely skewed distribution. The Type I error rates for MAD_n, T_n and LMS_n are also within the robust criterion under mildly skewed distribution with negative pairing especially for larger sample sizes. In general under this case, the Type I error rates of the conventional procedures are fell within the robust interval only for positive pairing for all type of distribution for Kruskal Wallis.

5.3 Implications

Our goal is to search for some alternative methods in testing the equality of location measures for skewed distributions. In this final chapter, we would like to share some of the advances that emerged from this study. Modifications made on the T_1 statistic successfully improved the performance of the statistic in terms of Type I error and power.

It is our impression that applied researchers would prefer a method that compared treatment performance across groups with a measure for the typical score which was based on as much as the original data as possible. T_1 statistic with asymmetric trimming *HQ1* will be the best choice for this purpose because when working with the T_1 with *HQ1* procedures, the researchers can work with the original data without having to worry about shape of the distribution.

These modified methods may serve as alternatives to some other robust statistical methods which are unable to handle either the problem of non-normality, variance heterogeneity or unbalanced design. This study may generate ideas for future research in robust methods simultaneously contributes to filling the gaps in the literature in this field.

5.4 Suggestion for Future Research

As stated in the earlier chapter, what is desired in this study is an inference procedure which in some sense performs almost as well as possible if the assumption is true, but does not perform much worse within a range of alternatives to the assumption. We have proved that some of the modified methods in this study were robust and performed remarkably well under violated assumptions, but when the assumption is true, such as under symmetric balanced design, the procedures failed to show robustness. Since we were able to improve the original methods by substituting the scale estimators used in the methods with some of the most robust scale estimators, this study should be continued with some other robust scale estimators in order to find solutions to the conservative Type I error rates and smaller power rates for the balanced design. Rousseeuw and Croux (1993) suggested plenty of alternatives that is worthy of consideration.

As accomplished in this study, by respectively substituting trimmed means and Winsorized variances (using trimmed mean) in place of usual means and variances, into T_1 statistic (Babu *et al.*, 1999), robustness to both nonnormality and variance heterogeneity can be achieved, even if the sample sizes are unequal.

To empirically determine how much trimming is really needed is difficult and not always obvious. However, if the data need to be trimmed, the trimming has to be done meticulously in order to avoid unnecessary trimming. This can be achieved by using the new trimming strategies with automatic trimming or asymmetric trimming proposed in this study. By using this approach, we do not trim a fixed amount of data. Othman *et al*. (2002) in their study reported from the theoretical considerations that when data are said to be skewed to the right, then in order to achieve robustness to nonnormality and greater sensitivity to detect effects, one should trim data just from the upper tail of the data distribution.

In this study, our main concern is about the trimming methods and how much trimming need to be done to avoid loss of important information when nonnormality and variance heterogeneity exist. In dealing with these problems, a few new trimming strategies were proposed. These trimming strategies will ensure that the aforementioned problems will be adequately addressed.

To improve the performance of the T_1 procedures, we used bootstrapping method to test the hypotheses. The reason of using bootstrapping method was due to the fact that the sampling distributions for the statistics used were intractable. Certainly, further research is required in arriving at the sampling distributions. Too much reliance on resampling techniques to come up with a pseudo sampling distribution goes against the grain of traditional mathematical statistics whereby a sampling distribution or an asymptotic sampling distribution is always preferable.

REFERENCES

Abdullah, S., Syed Yahaya, S.S. and Othman, A.R. (2008). A power investigation of Alexander Govern test using modified one-step m-estimator as a central tendency measure. *Proceedings of Joint Meeting of 4th. World Conference of IASC and 6th Conference of the Asian Regional Section of the IASC on Computational Statistics and Data Analysis*, Yokohama.

Alexander, R. A., & Govern. D. M. (1994). A new and simpler approximation for ANOVA under variance heterogeneity. *Journal of Educational Statistics*, **19**(2), 91 – 101.

Babu, J. G., & Padmanabhan, A. R. (1996). A robust test for omnibus alternatives. *Madan Puri Festschrift*, 319 – 327.

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

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

Brownlee, K.A. (1965). *Statistical theory and methodology in science and engineering* (*2nd Ed*.). New York: Wiley.

Carroll, R. (1982). Two examples of transformations when there are possible outliers. *Applied Statistics*, **31**, 149 - 152.

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

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. New York: Chapman & Hall.

Fenstad, G. U. (1983). A comparison between U and V tests in the Behrens-Fisher problem. *Biometrika*, **70**, 300 -302.

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

Granger, C.W. J. (1996), “Can we improve the perceived quality of economic forecasts?” *Journal of Applied Econometrics*, **11**, 455–473.

Gross, A. M. (1976). Confidence interval robustness with long-tailed symmetric distributions. *Journal of the American Statistical Association*, **71**, 409 – 416.

Hall, P. (1986). On the number of bootstrap simulations required to construct a confidence interval. *Annals of Statistics*, **14**, 1431 – 1452.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., & Stahel, W. A. (1986). *Robust statistics*. New York: Wiley.

Higazi, S. M. F., & Dayton, C. M. (1984). Comparing several experimental groups with a control in the multivariate case. *Communication in Statistics,* **13**, 227 – 241.

Hoel, P. G., Port, S. C., & Stone, C. J. (1971). *Introduction to statistical theory*. Boston: Houghton Mifflin.

Herron, R. E., & Hillis, S. L. (2000). The impact of the transcendental meditation program on government payments to physicians in Quebec: an update. *American Journal of Health Promotion,* **14**, 284-291.

Hertsgaard, D. (1979). Distribution of asymmetric trimmed means. *Communications in Statistics: Simulation and Computation,* **8**, 359 – 367.

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. 461 - 513.

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

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., Lix, L. M., & Kowalchuk, R. K. (1998). Multiple comparison procedures for trimmed means. *Psychological Methods,* **3**(1), 123 – 141.

Keselman, H. J., Kowalchuk, R. K., & Lix, L. M. (1998). Robust nonorthogonal analyses revisited: An update based on trimmed means. *Psychometrika,* **63**, 145 – 163.

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

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.

Keselman, H. J., Othman, A. R., Wilcox, R. R., & Fradette, K. (2004). The new and improved two-sample *t* test. *Psychological Science,* **15**, 47-51.

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.

Kulinskaya, E., & Dollinger, M. B. (2007). Robust weighted one-way ANOVA: Improved approximation and efficiency. *Journal of Statistical Planning and Inference,* **137,** 462 - 472.

Lee, H & Fung, K.Y. (1985). Behaviour of trimmed *F* and sine-wave *F* statistics in one-way ANOVA. *Sankhya: The Indian Journal of Statistics*, **47** (Series B), 186-201.

Levy, K. J. (1978). An empirical comparison of the *ANOVA F*-test with alternatives which are more robust against heterogeneity of variance. *Journal of Statistical Computation and Simulation*, **8**, 49 - 57.

Lix, M. L., 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.

Lix, L. M., & Keselman, H. J. (1998). To trim or not to trim: Tests of location equality under heteroscedasticity and nonnormality. *Educational and Psychological Measurement,* **58**(3), 409 – 429.

Luh, W. M., & Olejnik, S. (1990). *Two-stage sampling procedures for comparing means when population distributions are non-normal.* In Annual Meeting of the American Educational Research Association, 16 – 20 April, 1990, Boston, MA, pg 24.

Luh, W. M & Guo, J. H. (1999). A powerful transformation trimmed mean method for one-way fixed effects ANOVA models under non-normality and inequality of variances. *British Journal of Mathematical and Statistical Psychology*, **52**, 303-320.

Manly, B. F. J. (2007). *Randomization, Bootstrap and Monte Carlo Methods in Biology* (*3rd Ed*.). London: Chapman & Hall.

Mann, P. S. (2004). *Introductory statistics*. New York: Wiley.

Mason, R. D., Lind, D. A., & Marchal, W. G. (1999). *Statistical techniques in business and economics (10th Ed.).* Singapore: McGraw Hill.

Maxwell, S. E., & Delaney, H. D. (2004). Designing experiments and analyzing data (*2nd Ed.*). Mahwah, NJ: Erlbaum.

Myers, L. (1998). Comparability of the James' second-order approximation test and the Alexander and Govern A statistic for non-normal heteroscedastic data. *Journal of Statistical Computational Simulation, 60*, 207-222.

Oshima, T. C., & Algina, J. (1992). Type I error rates for James's second-order test and Wilcox's Hm test under heteroscedasticity and non-normality. *British Journal of Mathematical and Statistical Psychology,* **42**, 255-263.

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

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.

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

Rodrigues, P., & Rubia, A. (2006). Testing for structural breaks in variance with additive outliers and measurement errors. *Working Paper, Universidad de Alicante*.

Ronchetti, E. M. (2006). The historical development of robust statistics. In Proceedings of the 7th International Conference on Teaching Statistics (ICOTS-7). 2 – 7 July, 2006, Salvador, Brazil. International Statistical Institute: Working Cooperatively in Statistics Education.
[Retrieve 20 February 2008 from
http://www.stat.aucland.ac.nz/~iase/publications/17/3B1_RONC.pdf.].

Rousseeuw, P. J., & Leroy, A. M. (1987). *Robust regression and outlier detection.* 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.

Sawilowsky, S.S. (2002). A measure of relative efficiency for location of a single sample. *Journal of Modern Applied Statistical Methods,* **1**(1), 52 – 60.

Scheffe, H. (1959). *The analysis of variance*. New York: Wiley.

Schneider, P. J., & Penfield, D. A. (1997). Alexander and Govern's approximation: Providing an alternative to *ANOVA* under variance heterogeneity. *Journal of Experimental Education, 65*, 271-287.

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

Steven, J. (1990). *Intermediate statistics a modern approach*. NJ: Lawrence Erlbaum.

Syed Yahaya, S. S., Othman, A. R., & Keselman, H. J. (2004a). Testing the equality of location parameters for skewed distributions using S_1 with high breakdown robust scale estimators. *In M.Hubert, G.Pison, A. Struyf and S. Van Aelst (Eds.), Theory and Applications of Recent Robust Methods, Series: Statistics for Industry and Technology, Birkhauser, Basel. 319 – 328.*

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.

Tiku, M. L. (1980). Robustness of MML estimators based on censored samples and robust test statistics. *Journal of Statistical Planning and Inference*, **4**, 123 – 143.

Tiku, M. L. (1982). Robust statistics for testing equality of means and variances. *Communications in Statistics: Theory and Methods*, **11**, 2543 – 2558.

Thomas, G. E. (2000). Use of the bootstrap in robust estimation of location. *Journal of the Royal Statistical Society*. **49**(1), 63 – 77.

Tomarken, A. J., & Serlin, R. C. (1986). Comparison of *ANOVA* alternatives under variance heterogeneity and specific noncentrality structures. *Psychological Bulletin*, **99**, 90-99.

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

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

Wilcox, R. R. (1992). Why can methods for comparing means have relatively low power, and what can you do to correct the problem? *Psychological Science*, **1**(3), 104 – 105.

Wilcox, R. R. (1995). ANOVA: A paradigm for low power and misleading measures of effect size? *Review Of Educational Research*, **65**(1), 51 – 77.

Wilcox, R. R. (1996). A note on testing hypotheses about trimmed means. *Biometrical Journal*, **38**, 173-180.

Wilcox, R. R. (1998). The goal and strategies of robust methods. *British Journal of mathematical and Statistical Psychology*, **51**, 1 – 39.

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

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

Wilcox, R. R. (2005b). Comparing medians: An overview plus new results on dealing with heavy-tailed distributions. *The Journal of Experimental Education*, **73**(3), 249-263.

Wilcox, R. R., Charlin, V. L., & Thompson, K. L. (1986). New Monte Carlo results on the ANOVA *F*, *W* and *F** statistics. *Communications in Statistics, Simulation and Computation*, **15**(4), 933 – 943.

Wilcox, R. R., Keselman, H. J., Muska, J., & Cribbie, R. (2000). Repeated measures ANOVA: Some new results on comparing trimmed means and means. *British Journal of Mathematical and Statistical Psychology*, **53**, 69-82.

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

Wilcox, R. R., & Keselman, H. J. (2003). Modern robust data analysis methods: Measures of central tendency. *Psychological Methods,* **8**, 254-274.

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

Yuen, K. K. (1974). The two-sample trimmed *t* for unequal population variances. *Biometrika*, **61**, 165 – 170.