MODIFICATION OF S_1 STATISTIC WITH HODGES-LEHMANN AS THE CENTRAL TENDENCY MEASURE

LEE PING YIN

MASTER OF SCIENCE (STATISTICS)
UNIVERSITI UTARA MALAYSIA
2018

Awang Had Salleh
Graduate School
of Arts And Sciences

Universiti Utara Malaysia

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

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

LEE PING YIN

calon untuk Ijazah **MASTER OF SCIENCE (STATISTICS)**
(candidate for the degree of)

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

"MODIFICATION OF S1 STATISTIC WITH HODGES-LEHMANN AS THE CENTRAL TENDENCY MEASURE"

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

Bahawa tesis/disertasi tersebut boleh diterima dari segi bentuk serta kandungan dan meliputi bidang ilmu dengan memuaskan, sebagaimana yang ditunjukkan oleh calon dalam ujian lisan yang diadakan pada : ***08 Mei 2018.***
That the said thesis/dissertation is acceptable in form and content and displays a satisfactory knowledge of the field of study as demonstrated by the candidate through an oral examination held on:
May 08, 2018.

Pengerusi Viva: *(Chairman for VIVA)*	**Assoc. Prof. Dr. Nazihah Ahmad**	Tandatangan *(Signature)*
Pemeriksa Luar: *(External Examiner)*	**Assoc. Prof. Dr. Zulkifley Mohamed**	Tandatangan *(Signature)*
Pemeriksa Dalam: *(Internal Examiner)*	**Assoc. Prof. Dr. Zahayu Md Yusof**	Tandatangan *(Signature)*
Nama Penyelia/Penyelia-penyelia: *(Name of Supervisor/Supervisors)*	**Prof. Dr. Sharipah Soaad Syed Yahaya**	Tandatangan *(Signature)*
Nama Penyelia/Penyelia-penyelia: *(Name of Supervisor/Supervisors)*	**Dr. Nor Aishah Ahad**	Tandatangan *(Signature)*

Tarikh:
(Date) ***May 08 2018***

Permission to Use

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

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

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

Abstrak

Andaian kenormalan dan kehomogenan varians adalah merupakan perkara penting bagi prosedur parametrik seperti dalam pengujian kesamaan kecendurangan memusat. Sebarang ketidakpatuhan andaian tersebut boleh meningkatkan kadar Ralat Jenis I yang serius, yang akan mengakibatkan penolakan hipotesis nol yang tidak betul. Prosedur parametric seperti *ANOVA* dan ujian-*t* sangat bergantung pada andaian yang sukar ditemui dalam data sebenar. Sebaliknya, prosedur tak berparameter tidak bergantung pada taburan data tetapi prosedur tersebut kurang kuasanya. Untuk mengatasi isu yang dinyatakan, prosedur teguh adalah dicadangkan. Statistik S_1 adalah salah satu prosedur teguh yang menggunakan median sebagai parameter lokasi untuk menguji kesamaan kecenderungan memusat di antara kumpulan, dan ia membabitkan data asal tanpa perlu memangkas atau mentransformasi data untuk mencapai kenormalan. Kajian terdahulu terhadap S_1 menunjukkan kekurangan keteguhan dalam beberapa keadaan di bawah reka bentuk seimbang. Oleh itu, objektif kajian ini adalah menambahbaik statistik S_1 asal dengan menggantikan median kepada penganggar *Hodges-Lehmann*. Penggantian juga dilakukan terhadap penganggar skala menggunakan varians bagi penganggar *Hodges-Lehmann* serta beberapa penganggar skala teguh yang lain. Bagi memeriksa kekuatan dan kelemahan prosedur yang dicadangkan dalam mengawal Ralat Jenis I, beberapa pemboleh seperti jenis taburan, bilangan kumpulan, saiz kumpulan yang seimbang dan tidak seimbang, varians yang sama dan tidak sama, dan sifat pasangan telah dimanipulasikan. Hasil kajian menunjukkan kesemua prosedur yang dicadangkan adalah teguh merentasi semua keadaan bagi setiap kes kumpulan. Selain itu, tiga prosedur yang dicadangkan iaitu $S_1(MAD_n)$, $S_1(T_n)$ dan $S_1(S_n)$ menunjuk prestasi yang lebih baik berbanding prosedur S_1 asal di bawah taburan pencong yang ekstrem. Secara keseluruhan, prosedur yang dicadangkan menunjukkan keupayaannya mengawal peningkatan Ralat Jenis I. Oleh yang demikian, objektif kajian ini telah tercapai apabila tiga daripada prosedur yang dicadangkan menunjukkan peningkatan keteguhan di bawah taburan terpencong.

Katakunci: Statistik S_1, *Hodges-Lehmann*, penganggar skala teguh, ralat Jenis I, taburan terpesong

Abstract

Normality and variance homogeneity assumptions are usually the main concern of parametric procedures such as in testing the equality of central tendency measures. Violation of these assumptions can seriously inflate the Type I error rates, which will cause spurious rejection of null hypotheses. Parametric procedures such as ANOVA and *t*-test rely heavily on the assumptions which are hardly encountered in real data. Alternatively, nonparametric procedures do not rely on the distribution of the data, but the procedures are less powerful. In order to overcome the aforementioned issues, robust procedures are recommended. S_1 statistic is one of the robust procedures which uses median as the location parameter to test the equality of central tendency measures among groups, and it deals with the original data without having to trim or transform the data to attain normality. Previous works on S_1 showed lack of robustness in some of the conditions under balanced design. Hence, the objective of this study is to improve the original S_1 statistic by substituting median with Hodges-Lehmann estimator. The substitution was also done on the scale estimator using the variance of Hodges-Lehmann as well as several robust scale estimators. To examine the strengths and weaknesses of the proposed procedures, some variables like types of distributions, number of groups, balanced and unbalanced group sizes, equal and unequal variances, and the nature of pairings were manipulated. The findings show that all proposed procedures are robust across all conditions for every group case. Besides, three proposed procedures namely $S_1(MAD_n)$, $S_1(T_n)$ and $S_1(S_n)$ show better performance than the original S_1 procedure under extremely skewed distribution. Overall, the proposed procedures illustrate the ability in controlling the inflation of Type I error. Hence, the objective of this study has been achieved as the three proposed procedures show improvement in robustness under skewed distributions.

Keywords: S_1 statistic, Hodges-Lehmann, robust scale estimators, Type I error, skewed distributions.

Acknowledgement

First of all, I would like to thank God for giving me the chance to complete the thesis which I have spent five years in studying Master of Sciences (Statistics) as a part time student. This is truly a blessing to me. Besides, I would like to extend my appreciation to my supervisor, Associate Professor Dr. Sharipah Soaad Syed Yahya and co-supervisor, Dr Aishah Ahad who have given continuous guidance, patience and support to me. They always be there for me whenever I face difficulty in my journey of writing my thesis and generating data using statistical computer software. I appreciate their help so much. In addition, I would like to thank Universiti Utara Malaysia (UUM) too for approving my master study's application and few staffs in Awang Had Salleh Graduate School who assisted me in the process of submission.

I am deeply grateful to my family, my fiance, Chan Jin Swan and my best undergraduate roommate, Nurull Salmi Md Dazali, my Indonesia friend, Fera, my colleague, Wong Sock Leng and my mentor, Wern Lu who encourage me throughout this study by giving infinite motivation. Due to all the support I get, I manage to complete my study. I would like to give my grateful appreciations to all of them.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

ANOVA	Analysis of Variance
HL	Hodges-Lehmann Estimator
MAD_n	Median Absolute Deviation about the Median
SAS	Statistical Analysis Software
SAS/IML	Statistical Analysis Software/Interactive Matrix Language
SPSS	Statistical Package for Social Science

CHAPTER ONE
INTRODUCTION

1.1 Introduction

In most research, hypothesis testing has been used as a method of decision making with the help of primary and secondary data that can be obtained from sources such as observations, experiments, journals, articles, reference books and many other sources. The researchers are required to identify the statement of null hypothesis which is usually corresponds to a situation of equality or "no difference" and it is assumed as true hypothesis until receiving an evidence that shows otherwise. Alternative hypothesis is known as the negation of null hypothesis (Sullivan, 2004). Due to the statistical nature of a test, two types of error are determined, Type I error and Type II error. Type I error occurred in the situation where by the null hypothesis is rejected when it is true. In contrast, Type II error existed when the null hypothesis is failed to reject when it is false. There is an inverse relationship between the two errors such that an increase in Type I error will decrease Type II error and vice versa. Furthermore, when Type II error increases, the statistical power of a test will decrease, causing less detection of a test effect. Thus, these two errors need to be in control. A good statistical procedure should be able to control the errors. However, working with Type I error is easier than Type II error as the earlier is usually set in advance by the researcher while the latter is harder to know as it requires estimating the distribution of the alternative hypothesis (Ramsey, 2001).

In order to achieve a good test, we need an appropriate procedure which is able to control Type I error rate and increase the power at the same time. We do not want to lose power, and we do not want to inflate the Type I error rate too. There are several statistical procedures for testing the equality of location measures or locating treatment effects across groups by simultaneously controlling Type I error and improving power of the procedures in detecting the treatment effects have been studied in recent years.

Parametric procedures are widely used by researchers in many fields to test the equality of the location parameters due to precision and easy to compute. However, these procedures rely heavily on assumption of normality. For further understanding, the example of Analysis of Variance (ANOVA) and its disadvantages will be referred to with regards to violation of assumptions. ANOVA is one of the popular parametric statistical procedures which used to analyse the difference between the means for more groups in one-way independent group design. Independence of observations, normality and equality of variance (homoscedasticity) are the basic assumptions when applying this procedure. Nevertheless, violation of normality and homoscedasticity assumptions always occur in practice. These problems have degenerated the properties of Type I error and reduce the power of a test in detecting the treatment effect. When the underlying distribution has heavy tails such as symmetric heavy-tailed and skewed heavy-tail distributions, the standard error of the mean $\left(\sigma^2/n\right)$ can become seriously inflated and also reduce the power of test (Wilcox and Keselman, 2002).

Thus, nonparametric statistics occurred as a field of research and several procedures turn to be very famous in applications. Nonparametric procedures do not rely on any data belonging to any particular distribution. They are sometimes known as distribution free procedures and can be used on data such as residents' favourite TV programmes by rating them with the scale of 1 to 10 which 1 refers to the least favour and 10 is for the most favour. Making few assumptions about the data is the basic principle of nonparametric procedures and its applicability is much wider and more robust compared to parametric procedures in most cases. Nonparametric procedures are easier to apply in most cases even though the uses of parametric procedures are justified. Nevertheless, nonparametric procedures are less powerful and a larger sample size with the same degree of confidence is required in order to reject a false hypothesis (Gibbons and Chakraborti, 2003). Under this circumstance, the use of parametric and nonparametric procedures are not advisable. To overcome the problem, robust statistical procedures are used as the alternatives.

Huber (1964) and Hampel (1974) established a complete theory of robust statistics, which basically centered on parametric models. Robust statistical procedures generally are not unduly affected by departures from the model assumptions. Besides, construction of the statistical procedures are still reliable and practically efficient in a neighbourhood of the model concerned (Ronchetti, 2006). Normality, independence and homoscedasticity are among the classical assumptions that hardly fulfilled in practice. Any violation of these assumptions will lead to biased results when tests are conducted. A definition given by Hampel, Ronchetti, Rousseeuw and Stahel (1986) states that, "In a broad informal sense, robust statistics is a body of knowledge, partly formalised into "theories of robustness", relating to deviations from idealised assumptions in statistics."

The violation of normality assumption is among the most frequently discussed issue in robust statistics. This violation can reduce the power to a lower stage when the means of two or more groups are compared (Wilcox and Keselman, 2003). Refer to Md Yusof, Othman and Syed Yahaya (2010), in the study of robust statistics, Huber (1981), Staudte and Sheather (1990) and Wilcox (1997) considered robust measures of location like trimmed means or medians as the alternative solutions for the usual least squares estimator. According to Syed Yahaya, Othman and Keselman (2006), other studies had also proved that Type I error from the test of treatment effects can be well controlled through these measures of location (Othman, Keselman, Padmanabhan, Wilcox, and Fradette, 2004). A certain percentage of the smallest and largest observations are removed and averaging the remaining values is known as trimmed mean. The percentage of trimmed mean is fixed in advance to ease in analysing data. For example, 10% trimming is referred to 10% of the smallest observations and 10% of the greatest observations are trimmed. If there are 10 observations with the least value of 6 and greatest value of 35, 10% trimming is referred as removing the values of 6 and 35 follows by computing the average of the remaining observations. Yuen (1974) had found that there were some advantages on trimmed means for two groups case. Similar results on trimmed means for more than two groups case were established by Lix and Keselman (1998) and researchers were reminded that non-normality of one's data should not automatically signal the adoption of trimmed means and robust test statistics. Researchers should take serious consideration under such circumstances about the reasons of non-normality and also to examine the method of collecting data, measurements instruments and the process of generating data. Before applying trimmed mean in any research, researchers have to decide the percentage of trimming because this may cause the losing of some important information especially when the number of trimming has been decided prior to data analysis.

Besides trimmed mean, sample median, that is the midpoint in a set of observations is also known as one of the common robust estimators especially for sufficiently heavy-tailed distributions (Wilcox, 2012). It can endure large proportion of worst observations without breaking down completely since it has been characterised by the highest breakdown point (0.5). Refer to Donoho and Huber (1983), breakdown point is roughly the smallest amount of contamination that may cause an estimator to take on arbitrarily large aberrant values. This characteristic is very helpful in understanding the robustness properties of estimators. If there are n observations and let a minority of them $[(n-1)/2n]$ reach infinity leaving the rest fixed, then the median stays with the majority. Therefore, the breakdown point of median in finite sample is $[(n-1)/2n]$ and the asymptotic breakdown point is $1/2$ (0.5). According to Huber (1981), for an ideal parametric model, the estimator and statistical testing of central tendency measures are always misleading by a small number of extreme values on the data sets due to their lower breakdown points. Let take sample mean as the example. Given that the observations of $X_1,...,X_n$ and the formula for mean is as below:

$$\frac{x_1 + x_2 + ... + x_n}{n} \tag{1.1}$$

If the *n*th observation approaches infinity, the sample mean will fall to infinity too. This explains that the sample mean will be ruined even with one gross outlier. The breakdown point of sample mean in finite sample is $1/n$ and the asymptotic breakdown point is zero. This means that there is only $1/n$ sample breakdown point of sample mean when *n* approaches infinity.

The following scenario is one of the examples to demonstrate the effectiveness of a robust statistics. Given that there are five measurements of a concentration, 5.59, 5.66, 5.63, 5.57 and 5.60. Normally we will calculate the sample mean for estimating its true value. The usual average of these five numbers is 5.61. Another estimator that can be used is sample median and it yields 5.60. In this example, the values of mean and median are close to each other. Let us now suppose that one of the measurements is recorded wrongly such that the data is recorded as 5.59, 5.66, 5.63, 55.7 and 5.60. This situation often happens in research due to data entry error. There is also a possibility that the outlying observation is incorrect or it belongs to other population. Under this circumstance, the mean becomes 15.64. In contrast, the value of the median is 5.63, which is still reasonable despite the error. The median is changed but it does not become arbitrarily bad as mean. However, median is also known as trimmed mean with 50% since 50% of the largest observations and 50% of the smallest observations are removed. This may also cause a losing of some important information.

Yi and He (2009) had done a research for longitudinal data with dropouts using median regression model. As discussed in Morgenthaler (1992), modeling the study was partly inspired by longitudinal data arising from a controlled trial of HIV disease. One of the main objectives was to examine the treatment effect of Zidovudine on growing CD4+ cell counts. 892 adults were randomised to a treatment group and they were tracked longitudinally. At weeks 8, 16, 32 and 48, the measurements were collected. They tested the data using median regression model as well as mean regression model for comparison purpose. Based on the results obtained, the proposed median regression procedure performed well for a range of data with different distributions. However, mean regression approach relied on the distributional shapes. No doubt, it provided accurate results for normal distribution but it may unable to give reliable results for a data with other distributions.

1.2 Problem Statement

The letdown of parametric procedures in dealing with non-normal data and a few challenging assumptions obliged the users of statistics to opt for alternative procedures in nonparametric as well as robust procedures. However, nonparametric procedures also have their drawbacks especially in terms of losing information due to ranking process turned the users to a more reliable procedure in robust statistics.

A good robust estimator combined with good statistical procedures might be able to solve some of the typical problems encountered by the users of statistics. One of such procedures is S_1 statistic which was proposed by Babu, Padmanabhan and Puri (1999) when the distributions are skewed. This procedure comes with the purpose of measuring treatment effects across two and more than two groups by using median as the location parameter. As explained in previous section, median is the midpoint in a set of observations which can endure large proportion of worst observations without breaking down completely by its highest breakdown point (0.5). Therefore, S_1 statistic can deal with the original data without having to trim or transform it to attain symmetry.

Indeed, real data rarely fulfill the assumption of normality. As Reed (1998) quoted: "Nearly all real data are discrete in nature therefore the theory suggests that they cannot be normal". According to Maxwell and Delaney (2004), the main disadvantage of data transformation is the interpretation of results may be less than clear because researchers are working in a metric other than the original variable. Besides, finding a transformation which will deal with asymmetry and variance heterogeneity simultaneously is difficult (Keselman, Wilcox, Lix, Algina and Fradette, 2007). On the other hand, trimming also may cause the losing of some important information of the data as explained in Section 1.1.

Othman *et al*. (2004) modified S_1 statistic by replacing the standard errors of the sample medians with asymptotic variances by referring to Hall and Sheather's (1988) work on sample medians. For comparison purposes, the proposed procedure and the original S_1 statistic were tested under the condition of non-normality and variance heterogeneity for two groups and four groups cases. The finding showed that the proposed procedure generated slightly closer Type I error rates to nominal level of 0.05 than the original S_1 statistic for four groups case. However, these error rates were lower (deviated further away from the nominal value) than the Type I error rates produced under two groups case and were considered non robust. In addition, the proposed procedures failed to show better control of Type I error compared to the original S_1 statistic for two groups case.

Syed Yahaya (2005) proposed a study on testing the equality of location parameter in one-way independent group design when the distributions were skewed. S_1 statistic was selected as the procedure of the study and being modified by replacing the default scale estimator, $\hat{\omega}$ with four robust scale estimators, MAD_n, Q_n, S_n and T_n. MAD_n is known as the famous robust scale estimator with its highest breakdown point and having the capability of maintaining the robustness of procedures. The findings proved that three out of four proposed S_1 procedures using MAD_n, S_n and T_n as the scale estimators had good control of Type I error rates compared to the original S_1 statistic under extremely skewed distributions for two and four groups cases. However, the previous work on S_1 observed that most of the conditions under four groups case were non robust especially under the influence of extremely skewed distribution. Other than the issue of robustness, the proposed procedures generated conservative Type I error rates (below 0.025 level) in most conditions for both group designs. Thus, by using different types of location estimators, while maintaining the four robust scale estimators used in the previous study, we expect to have some significance improvement of S_1 statistic in terms of controlling Type I error for skewed distributions.

1.3 Objective(s) of the Study

The purpose of this study is to improve the original S_1 statistic for testing the equality of central tendency measures in one-way independent group design under skewed distributions. In order to achieve this goal, the objectives below are required to be accomplished.

i. To modify the S_1 statistic by replacing the median with the Hodges-Lehmann.

ii. To evaluate the modified procedures with simulated data.

iii. To compare the modified procedures against some parametric and nonparametric procedures in terms of the empirical Type I error rates.

iv. To compare the modified procedures against some parametric and nonparametric procedures on real data.

v. To identify the best procedures.

1.4 Significance of Study

This study will significantly contribute to the body of knowledge in statistical procedure especially experimental design which usually attached with strict assumptions such as normality and variance homogeneity to achieve reliable results. The proposed procedure gives some flexibility to the users for testing treatment effects between groups, unlike parametric procedures such as *t*-test and ANOVA when the violation of assumptions exists. This flexibility should be a welcome feature for industries as they are always depend on easy to compute, fast and trustworthy statistical procedure to be employed because of the challenge of obtaining real data which is often required to fulfill the assumption of normality. Even to those users of statistical procedures who are not constantly aware of or do not pay attention on the assumptions, this type of procedure will suit them well due to no additional work is needed to perform before applying it. Furthermore, it doesn't jeopardise the results.

1.5 Organisation of the Thesis

The background of parametric, nonparametric and robust statistics are mentioned briefly in this chapter. Besides, one of the robust procedures, S_1 statistic is introduced. Further information about this procedure and recommended scale estimators will be explained in Chapter 2. In addition, some commonly used parametric and nonparametric procedures will be reviewed too. Chapter 3 will show the employment of proposed procedure and manipulation of variables. In addition, this chapter will also describe the design specification of this study. Type I error rates of each procedure will be presented and analysed in Chapter 4. Lastly, Chapter 5 will be the final chapter of the thesis that includes conclusion and suggestions for further studies.

CHAPTER TWO
LITERATURE REVIEW

2.1 Introduction

Parametric and nonparametric statistical procedures are available in statistical inference or hypothesis testing. Parametric procedures rely on assumptions heavily such as normality and variance homogeneity with regards of the distributional shape in the underlying population and the location parameter of the distribution. Violation of these two assumptions are often the major practical problems that is encountered by researchers when using parametric procedures especially on testing the equality of location measures for two and more than two independent groups. Conversely, nonparametric procedures do not rely on the assumptions about the distributional shape from which the sample was drawn. However, nonparametric procedures are less powerful compared to parametric procedures. Besides, in order to reject a false hypothesis, a larger sample size with the same degree of confidence is required, but practically, smaller sample size is more preferable (Gibbons and Chakraborti, 2003). Under such circumstances, the use of parametric and nonparametric procedures are not the better choice. Hence, to overcome the problem, robust statistical procedures are used as the alternatives.

Eighteenth century was the beginning of robust statistics when the first rules of outliers' rejection were developed. In nineteenth century, these rules were formalized and implemented for estimating mean. This was followed by the development of estimators that down weight outliers. First half of the twentieth century was the period where robustness of statistical testing being considered. The need for robust procedures was demonstrated by Box (1953) and Tukey (1960) (Stigler, 2010). Their research could be seen as the discovery of robust statistics. A few years later, Huber (1964) and Hampel (1974) established a complete theory of robust statistics, which basically centered on parametric models. As mentioned in Chapter One, robust procedures generally are not unduly affected by departures from the model assumptions. The following sections will discuss about the parametric, nonparametric and robust procedures for two and more than two groups cases that are frequently used and available in most statistical software.

2.2 Two-group Case

Suppose that $x_{1i},\ldots,x_{ni}$ and $x_{1j},\ldots,x_{nj}$ are the independent random samples from two populations which have continuous distribution function of $F(x_i)$ and $F(x_j)$ respectively. Assuming the two populations' variances are equal, there are few procedures available for testing the equality of the location parameter. Each of the commonly employed procedures from the parametric, nonparametric and robust approaches for testing the equality of central tendency measure will be discussed in the following sections in regards to their applications when violation of assumptions occur in the data. The procedures for the two group case are *t*-test, Mann-Whitney and S_1 Statistic.

2.2.1 *t*-test

The *t*-test is frequently used in comparing means between two groups. The validity on drawing the accurate inferences maybe weakened if the following assumptions of *t*-test are not met.

i. Two samples are independent.

ii. The populations follow normal probability distribution.

iii. The variances of both populations are equal.

The violation of any assumption above will increase Type I error rates and also reduce the power of the procedures at the same time (Maxwell and Delaney, 2004).

When sample sizes between two groups are equal, *t*-test can be calculated as:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\tilde{s}_1^2}{n_1} + \frac{\tilde{s}_2^2}{n_2}}} \qquad (2.1)$$

where

$\bar{x}_1$ = sample mean of group 1

$\bar{x}_2$ = sample mean of group 2

$\tilde{s}_1^2$ = estimated population variance of group 1

$\tilde{s}_2^2$ = estimated population variance of group 2

n_1 = sample size of group 1

n_2 = sample size of group 2

For unequal sample sizes, *t*-test can be computed as:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\left[\frac{(n_1 - 1)\tilde{s}_1^2 + (n_2 - 1)\tilde{s}_2^2}{n_1 + n_2 - 2}\right]\left[\frac{1}{n_1} + \frac{1}{n_2}\right]}} \qquad (2.2)$$

The degree of freedom which is used for significance testing is $2n - 2$ for equation (2.1) and $n_1 + n_2 - 2$ for equation (2.2).

Kang and Harring (2012) did a study about the impact of non-normality, effect size and sample size on two groups case for independent samples *t*-test. Monte Carlo simulation with 1000 replications was used to investigate on the robustness and Type I error on two equal group sizes under non-normal distributions. The five proposed group sample sizes were $n_1 = n_2 = 8$, $n_1 = n_2 = 15$, $n_1 = n_2 = 30$, $n_1 = n_2 = 60$ and $n_1 = n_2 = 120$. In regards to their findings, when the distributions for both groups were non-normal and had the same distributional shape, *t*-test managed to maintain its nominal Type I error rates ($\alpha = 0.05$) across different sample sizes. When two distributions were non-normal and had different shapes of distribution, Type I error was slightly inflated for the sample sizes of $n_1 = n_2 = 8$, $n_1 = n_2 = 15$ and $n_1 = n_2 = 30$. Yet, Type I error rates were able to be upheld at its nominal level as sample size increased.

Kellermann, Bellara, Gil, Nguyen, Kim, Chen and Kromrey (2013) did a research on variance heterogeneity and non-normality using SAS Proc Test which is an easy way of testing the equality of central tendency measure for two groups. The purpose was to discover *t*-test's performance under departure of normality and variance heterogeneity. The variables that were manipulated into several conditions consisted of the total sample size, ratio of sample size, effect size for mean difference, significance level for testing the treatment effect and the alpha level for testing homogeneity. As predicted, *t*-test was found to perform very well in controlling Type

I error when the variances were equal under equal or unequal sample sizes regardless of the tenability for the assumption of normality. No doubt, *t*-test emerged as the best procedure to test the difference of two independent means under this condition. However, their *t*-test could not adequately control Type I error when the group variances were not equal especially for unequal sample sizes. Besides, *t*-test showed a reduced in percentage on its statistical power under skewed distribution.

2.2.2 Mann-Whitney

Mann-Whitney test (Wilcoxon, 1945) is also known as Mann-Whitney U test or Wilcoxon-Mann-Whitney test. It is an alternative approach of *t*-test for comparing the difference in means between two groups when the assumption of normality and variance homogeneity are not met. Its effectiveness is quite similar as *t*-test on normal distribution (Sheskin, 2011). Refer to Gibbons and Chakraborti (2003), ranked data is used by Mann-Whitney for testing the central tendency measure by changing the actual numerical data to ranks in combined groups. The ranks obtained are then compared with the sums of ranks in two groups. The sampling distribution of Mann-Whitney test is approximately normal when both sample sizes are more than 10 and z test is used for statistical inferences. It is defined as:

$$z = \frac{T - \mu_T}{\sigma_T} \tag{2.3}$$

where

T = total of the ranks for the observations from the sample

μ_T = mean of the sampling distribution of T

σ_T = standard deviation of the sampling distribution of T

$$\mu_T = \frac{n_1(n_1 + n_2 + 1)}{2} \quad (2.4)$$

$$\sigma_T = \sqrt{\frac{n_1 n_2}{12}(n_1 + n_2 + 1)} \quad (2.5)$$

Two independent samples are required and the population distributions of both samples should be equal with the exception of the central measure for the purpose of drawing valid inferences from Mann-Whitney procedure (Ott & Longnecker, 2010). Winter and Dodou (2010) studied about the comparison between Mann-Whitney and *t*-test in terms of Type I and Type II error rates for five-point Likert items. In the study, pairs of samples were drawn from fourteen diverse Likert population distributions which were considered as the representative of the possible distributions that might appear in the real Likert item data. There were ten thousand random samples selected for each of the 98 combinations of distributions. For equal sample sizes, the simulations were conducted with $m = n = 10$, $m = n = 30$ and $m = n = 200$. For unequal sample sizes, $m = 20$, $n = 5$, and $m = 10$, $n = 100$, were used. The finding showed that both procedures have same power in general except for peaked, skewed, or multimodal distributions which Mann-Whitney produced better power than *t*-test. On the other hand, when $m = 20$, $n = 5$, Mann-Whitney was unable to show robustness and good control of Type I error on one Likert item with its error rate of 0.077 at nominal significance level of 0.05. For the same Likert item and unequal sample sizes ($m = 20$, $n = 5$), *t*-test also couldn't control Type I error well due to its error rate of 0.074.

Nachar (2008) used Mann-Whitney as the procedure for assessing whether two independent samples came from the same distribution. The investigation was carried out on two different groups of individuals with social phobia. One group was referred to those who received behavioral therapy and another group was for the people who accepted the combined therapy of behavioral and the antibiotic with n observations each. Since both groups showed a reduction in the number of symptoms of social phobia after each therapy, the number of these symptoms was then measured and tested under the sample sizes of 3, 4, 5, 6, 7 and 8. The finding showed that Mann-Whitney was more powerful and had better control of Type I error than t-test when the sample size was small. However, Type I error was amplified when there was a violation of variance homogeneity.

When both populations are not normally distributed and skewed into the same direction, Mann-Whitney yielded higher power rates compared to t-test. However, if the two groups shows different shapes of distributions, it might not be valid due to the increased in Type I error rates especially when the sample size is large (Kang and Harrings, 2012). To alleviate the problems that typically occur in t-test and Mann-Whitney procedures, the alternative is via robust procedures.

2.2.3 S_1 Statistic

One of the robust procedures that can be employed to test the equality of location parameter for two and more than two groups is S_1 statistic that was proposed by Babu, *et al* (1999). As explained in the previous chapter, S_1 statistic uses median as location parameter to measure the treatment effects across groups. Median is referred to as the middle value of a data set. When using this procedure, the original data can be used without going through the process of transformation or trimming. For example, if a data set contains values of 3, 4, 4, 5, and 8, the mean value is 4.8. If the value of 8 is entered as 89, the sample mean will change to 21. However, under the same situation, the value of median still remains as 4 which shows that as a location measure, median is robust or not sensitive to extreme values. Hence, S_1 statistic which uses median as the location measure is recommended as an alternative robust procedure especially in dealing with skewed distributions.

To understand S_1 statistic, consider the problem of comparing central tendency measures under skewed distributions. Let $Y_{ij} = \left(Y_{1j}, Y_{2j}, ..., Y_{n_j j}\right)$ be a sample from an unknown distribution F_j and let M_j be the population median F_j: j = 1, 2,… J. For testing H_0: $M_1 = M_2 = \ldots = M_j$ versus H_1: $M_i \neq M_j$ for at least one pair of (i, j), the S_1 statistic is defined as:

$$S_1 = \sum_{1 \le i < j \le J} \left| s_{ij} \right| \tag{2.6}$$

where

$$s_{ij} = \frac{(\hat{M}_i - \hat{M}_j)}{\sqrt{(\hat{\omega}_i + \hat{\omega}_j)}}; \tag{2.7}$$

$\hat{M}_i$ = the median of group i;

$\hat{M}_j$ = the median of group j;

$$\hat{\omega}_i = \frac{\omega_i}{n_i}; \hat{\omega}_j = \frac{\omega_j}{n_j}; \tag{2.8}$$

n_i = number of observations for group i;

n_j = number of observations for group j;

$$\omega_i = \left(\frac{1}{n_i} \sum_{i=1}^{I} \sum_{j=1}^{n_i} \left| X_{ij} - \hat{M}_i \right| \right)^2 \qquad (2.9)$$

$$\omega_j = \left(\frac{1}{n_j} \sum_{j=1}^{J} \sum_{i=1}^{n_j} \left| Y_{ij} - \hat{M}_j \right| \right)^2 \qquad (2.10)$$

S_1 in formula (2.6) is referred to the total of all possible differences between sample medians from the J distributions divided by square root of the sum of sample standard errors of sample medians, $\hat{\omega}$. Hence, the number of possible differences is similar to $J(J-1)/2$ if there are J distributions.

Syed Yahaya (2005) applied S_1 statistic in the study of "Robust statistical procedures for testing the equality of central tendency parameters under skewed distributions" for two groups and more than two groups. Four robust scale estimators, MAD_n, T_n, S_n and Q_n were used to replace the default scale estimator of S_1 statistic. Based on Bradley's liberal robust criterion, S_1 statistic with MAD_n and T_n were considered robust across all distributions for both group designs in two groups case. For four groups case, these two procedures were considered robust for normal and mildly skewed distributions under unbalanced group design. Besides, they did not show worse performance compared to the original S_1 statistic for symmetric and mildly skewed distributions. On the other hand, S_1 statistic with S_n also did better than original S_1 statistic under extreme conditions. However, the procedure of S_1 statistic with Q_n provided a very conservative and non-robust value for two groups and more than two groups. In addition, all four procedures were not considered robust across all three distributional shapes for balanced group design under four groups case. The generated Type I error rates were conservative in all conditions for the same group design and group case. Same issues of not robust and generating conservative Type I error rates happened on extremely skewed distribution for unbalanced group design under four groups case as well. Therefore, a suggestion for further modification on the location and scale estimators was given by the author in order to improve the performance of the S_1 statistic in terms of controlling Type I error.

2.3 More than Two Groups

There are few popular parametric and nonparametric procedures readily available for testing more than two groups. One of the commonly used parametric procedure is Analysis of Variance (ANOVA) when the populations are normally distributed. If the assumption of normality is violated, nonparametric procedure such as Kruskal-Wallis is used as the alternative procedure. A robust procedure like S_1 statistic also can be applied on more than two groups case. The following sections will further explain on the aforementioned procedures.

2.3.1 Analysis of Variance (ANOVA)

Analysis of Variance (ANOVA) is known as a parametric procedure for testing the equality of two or more than two population means. There are some assumptions that have to be met. The populations are normally distributed, independent and have equal variances. However, these assumptions are hardly fulfilled in practice. Similar to *t*-test, the violation of assumptions give the impacts on controlling Type I error and reducing the power of the test at the same time.

The test statistic for ANOVA is based on F distribution which is a continuous theoretical probability distribution which the F value will often fall within the range $0 \le F \le \infty$ (Sheskin, 2011). It can be computed as:

$$F = \frac{MSB}{MSW} \tag{2.11}$$

where

$$MSB = \frac{n_1(\bar{x}_1 - \bar{x})^2 + n_2(\bar{x}_2 - \bar{x})^2 + \ldots + n_k(\bar{x}_k - \bar{x})^2}{k - 1} \tag{2.12}$$

n_1 = sample size of population 1

n_2 = sample size of population 2, and so on

k = number of populations

$\bar{x}_1$ = sample mean of population 1

$\bar{x}_2$ = sample mean of population 2, and so on

$\bar{x}$ = sample mean of the combined data set

$$MSB = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \ldots + (n_k - 1)s_k^2}{n - k} \tag{2.13}$$

s_1^2 = sample variance of population 1

s_2^2 = sample variance of population 2, and so on

n = total number of observations of the combined data set

James (1951) and Welch (1951) recommended the estimation of the inverses of the variances of the respective sample means could be explained by weighting the terms in the sum of squares for larger sample sizes. Referred to Syed Yahaya, Md Yusof and Abdullah (2011), although ANOVA is generally known to be robust to small departure from normality, the extent of this departure is unknown unless the sample size is large enough to ensure the normality. Brownie and Boss (1994) studied on the robustness of ANOVA when the number of treatments is large using agricultural screening trials which often in a blocked design with limited replication (number of blocks). The objective of the research was to identify the existence of real differences between treatments. Besides, they also wanted to determine whether the procedure could provide good performance on Type I error and have good power at the same time. Therefore, the null hypothesis, H_0, of the study was “no differences between treatments”. Based on their findings, ANOVA was considered robust only for large number of blocks under H_0 although earlier statisticians had proved that both One and Two Ways ANOVA procedures were robust to non-normality if either the number of blocks or treatments was large (Scheffe, 1959). Furthermore, the procedure couldn’t provide good power for the data with frequent extreme values.

Lix, Keselman and Keselman (1996) suggested two approaches that might be considered by the researchers when the violations of assumptions were taking place. The first approach was applying a transformation on the data and proceed with ANOVA. However, there are some limitations of transformations. The researchers may face the difficulty in interpreting the outcomes since the conclusions have to be made based on the transformed scores instead of the original observations. Besides, according to Oshima and Algina (1992), there are several transformations which can be employed on a data set that depends on the specific type and degree of assumption violation that occur. This may not always be the simple solution for researchers. Selecting an alternative statistical procedure to ANOVA which is not sensitive to the assumptions such as nonparametric or robust procedure is the second approach. A nonparametric or robust procedure should be able to produce Type I error rate that is close to the nominal significance level, α, without having to concern much on the violation of assumptions. In addition, the alternative will also maintain the actual statistical power near to theoretical power (Lix, Keselman & Keselman, 1996). Further explanation of nonparametric and robust procedures will be shown in the following sections.

2.3.2 Kruskal-Wallis

Kruskal-Wallis is the well-known alternative procedure to one-way ANOVA for comparing the difference of central tendency using ranked data for at least three groups when the samples fail to meet the assumption of normality. It is a nonparametric procedure and also known as the extension of Mann-Whitney-Wilcoxon test to a design that involves more than two groups. If the result of Kruskal-Wallis is significant, it indicates that there is a significant difference across groups in the set of k groups.

All the observations in each sample that is from the different distributions are ranked from the smallest to the largest values. If there are two or more than two observations with the same value, mean of the ranks for tied values is computed. A computational formula for Kruskal-Wallis is shown below:

$$H = \frac{12}{N(N+1)}\left[\frac{R_1^2}{n_1} + \frac{R_2^2}{n_2} + \ldots + \frac{R_k^2}{n_k}\right] - 3(N+1) \qquad (2.14)$$

where

R_1^2 = sum of the ranks squared of group 1

R_2^2 = sum of the ranks squared of group 2 and so on

n_1 = number of observations in group 1

n_2 = number of observations in group 2

N = total number of observations

k = number of populations being compared

Khan and Rayner (2003) studied the robustness to non-normality of common tests, namely ANOVA and Kruskal-Wallis for more than two groups' location problem. The power functions for both procedures under several conditions were generated using simulation with *g* (for skewness) and *k* (for kurtosis) distribution which was suggested by MacGillivray and Cannon (2002). Based on the results obtained, Kruskal-Wallis performed better than ANOVA when sample sizes were large and kurtosis was high. The increase in sample size would radically improve Kruskal-Wallis's performance. However, Kruskal-Wallis did not seem to be an appropriate procedure for small sample sizes such as n < 5 especially when normality was violated.

Kruskal-Wallis procedure must be used with caution. It is similar to *F*-test which is sensitive to the occurrence of heterogeneous variances in equal and unequal sample sizes (Kruskal and Wallis, 1952). Lix. *et al.* (1996) did a research on a quantitative review of alternatives to One-Way ANOVA, Kruskal-Wallis test was chosen as the alternative procedure. Similar to ANOVA, Kruskal-Wallis should be sensitive to the violation of variance homogeneity under both balanced and unbalanced designs.

However, the outcome of the research did not support the statement (sensitivity due to heterogeneous variances) due to the generated data. In their study, almost 90% of the balanced design data (equal group size and equal group variance) were generated whereas there were only a small percentage of unbalanced design data (unequal group size and unequal group variance) were formed. Therefore, it was difficult to create clear guidelines about the use of Kruskal-Wallis test under variance heteroscedasticity. Other than this, the procedure showed good control for non-normality data. Nevertheless, it did not perform well when the non-normality was found under the populations with different distributions.

The choice of estimators is crucial in controlling Type I error rate and maintaining the power of statistical procedure. Due to the violations of normality and variance homogeneity, robust estimators have received a lot of attention in the literature by wide spread list in review articles (Huber, 1972; Hogg, 1974; Dixon and Yuen, 1974). Most of the robust estimators have been established and assessed for symmetric distributions with varying degrees of heavy tailed. According to Wilcox and Keselman (2003), using robust estimators can have significantly more power when the distributions of populations are differ in skewness or have unequal variances. In addition, they show better control of Type I error. Classical parametric procedures can be considered as robust by replacing the central tendency measures with robust estimators. Hence, in order to achieve the goal of this study, Hodges-Lehmann was chosen as the robust estimate of location.

2.4 Hodges-Lehmann Estimator

There was a serious rejection to classical statistical procedures based on linear models or non-normality is their susceptibility to gross errors for example heavier tails than the normal distribution. This issue had overcome successfully by nonparametric procedures such as Mann-Whitney or Kruskal-Wallis. The statistical power of these procedures are more robust against gross errors that parametric procedures like *t*-test and ANOVA. Modifying the classical location estimators through removal or winsorisation of outlying observations was a challenge to researchers. Hence, Hodges and Lehmann (1963) had introduced a different approach, Hodges-Lehmann estimator to these problems.

Hodges-Lehmann is a robust location estimator that derived from rank test statistics like Wilcoxon or normal scores statistics which were providing robust power successfully for the corresponding testing problems (Hodges and Lehmann, 1963). According to Boos (1982), it is a consistent and median-unbiased estimator of population mean under symmetric distribution. For skewed distribution, it estimates the "pseudo-median" which is related to population median closely. Furthermore, it is well known for having excellent robustness and efficiency properties under the usual assumption of symmetry.

Hodges-Lehmann estimator can be computed in a quick way. Let $X_1, ..., X_n$ be the sample from a continuous distribution $F(x) = F_0(x-\theta)/\sigma$. It is given as:

$$\hat{\theta} = median\left\{\frac{X_i + X_j}{2}, 1 \le i \le j \le n\right\} \quad (2.15)$$

The above formula is used to calculate for each group which is similar to a group mean.

Bickel (1965) studied on several robust estimates of location such as trimmed mean, winsorised mean, Hodges-Lehmann estimator and maximum likelihood estimator. By comparing to the trimmed mean, the finding suggested that Hodges-Lehmann estimator was to be preferred in any condition where the degree of contamination and shape of distribution is not known with great precision provided the computations involved are excessive. Besides, the conclusion of the study showed that all selected robust estimators, except winsorised mean behaved satisfactorily when compared to mean and Hodges-Lehmann estimator seemed to be the safest among the estimators.

Boos and Monahan (1986) proposed a procedure of incorporating prior information by replacing the likelihood in Bayes's formula with a bootstrap estimate of the sampling density of a robust estimate of location such as trimmed mean, sample mean, sample median and Hodges-Lehmann estimator. Laplace, Uniform and Student's *t* (3 *df*) were the three alternative error distributions that was considered in the study with the scale of unit variance each. Based on the results obtained, all four robust estimators showed a substantial improvement for Laplace and *t* distributions. Furthermore, Hodges-Lehmann estimator and trimmed mean provided the best results for *t* distribution by reducing the mean squared error of the posterior mean by approximately 60%.

2.5 Scale Estimators

In statistics, scale estimators are used to quantify the statistical dispersion in data sets. The sample standard deviation which is the common measures of scale is easily influenced by extreme value. The choice of scale measure in a test statistic is vital as this measure greatly influenced the result of the test. Othman *et al*. (2004) had tried to modify S_1 statistic by replacing standard error of sample median with asymptotic variances. However, this modification was not successful as Type I error was unable to be controlled at nominal level. Syed Yahaya, Othman and Keselman (2004) then continued working on this procedure by substituting four robust scale estimators such as MAD_n, T_n, S_n and Q_n in place of asymptotic variances. The substitution effectively controlled the Type I error under normal to moderately skewed distribution but failed to do so under extremely skewed distribution. T_n, S_n and Q_n were the robust scale estimators that introduced by Rousseeuw and Croux (1993). All the four estimators were selected according to their high breakdown point and bounded influence function, which were the two important characteristics of a scale estimator.

In the next section, random sample from any distribution will be represented as $X = (x_1, x_2, ..., x_n)$ and $med_i x_i$ refers to sample median for group *i*.

2.5.1 MAD_n

Median absolute deviation about the median, MAD_n is a frequently used robust scale estimator by researchers due to its best possible breakdown point of 50% which is doubled a number of interquartile range. Besides, its bounded influence function is the sharpest possible bound (Hampel, 1974). The formula is given by:

$$MAD_n = b\ med_i \left| x_i - med_j x_j \right| \tag{2.16}$$

Hampel (1974) was the first person who promoted MAD_n and he attributed it to Gauss. In the formula, the constant b is needed in order to make the estimator to remain consistent for the parameter of interest. MAD_n has a simple explicit formula and only need a little time for computation. Due to the benefits of MAD_n, Huber (1981) had concluded that MAD_n has developed as the single most useful robust scale estimator. However, there are some disadvantages of MAD_n. According to Rousseeuw and Croux (1993), it took a symmetric view on the dispersion because of the one first estimated the median and then attached the equal importance to positive and negative deviations from it. This situation did not seem to be a natural approach at asymmetric distributions which MAD_n was supposed to find the symmetric near the median that consisted 50% of the data.

2.5.2 S_n

Refer to Rousseeuw and Croux (1993), S_n is one of the alternative estimators for MAD_n that is used as initial or ancillary scale estimates in the similar way as MAD_n. It is also able to provide high efficiency and do not slanted towards symmetric distributions. It can be defined as:

$$S_n = c\ med_i \left\{ med_j \left| x_i - x_j \right| \right\} \tag{2.17}$$

The value of c is a constant factor and its default value is 1.1926. The notation, med_i is referred to the low median with the order statistic of rank $\left[(n+1)/2 \right]$ while med_j is meant for the high median with the order statistic of rank $h = (n/2) + 1$. This formula is quite similar to MAD_n but there is a slight difference between them. The operation of med_j was moved outside the absolute value.

S_n is always uniquely defined due to its explicit formula. One of the advantages of S_n that can overcome MAD_n's drawbacks is it does not require any location estimate of the data. S_n focuses on the typical distance between observations which is still under asymmetric distributions instead of measuring the distance between observation and the central value. Furthermore, S_n has the greatest possible breakdown point in finite sample by obtaining 58.23% efficiency which was better than MAD_n's 36.74% efficiency at Gaussian distributions (Rousseeuw and Croux, 1993).

2.5.3 Q_n

Q_n is the other alternative estimator that was suggested by Rousseeuw and Croux (1993). This estimator is defined as:

$$Q_n = d\left\{\left|x_i - x_j\right|; i < j\right\}_{(k)} \tag{2.18}$$

where d is a constant factor and

$$k = \binom{h}{2} \approx \binom{n}{2}\Big/4 \text{ and } h = (n/2)+1 \tag{2.19}$$

Other than its simple explicit formula that is suitable for asymmetric distributions and its high breakdown point of 50%, Q_n has a smooth bounded influence function that yields about 82% efficiency at Gaussian distributions, which is higher than S_n. Unfortunately, Q_n lost its efficiency in small samples.

2.5.4 T_n

Rousseeuw and Croux (1993) had proposed another simple explicit scale estimator with high breakdown point and suitable for asymmetric distribution, denoted as:

$$T_n = 1.3800 \frac{1}{h} \sum_{k=1}^{h} \left\{ \underset{j \neq i}{med} \left| x_i - x_j \right| \right\}_{(k)} \tag{2.20}$$

They proved that T_n has high breakdown point of 50%, a continuous influence function with 52% efficiency which was more efficient than MAD_n.

According to all the estimators' properties like breakdown point, bounded influence function and efficiency, these estimators were decided to be used as the scale estimator for S_1. Syed Yahaya (2005) proposed a study to seek for alternative procedures in testing the equality of location parameter in one-way independent group design when the distributions were skewed. In order to achieve the goal of the study, S_1 statistic was modified by replacing the default scale estimator, $\hat{\omega}$, with some robust scale estimators such as MAD_n, Q_n, S_n and T_n. According to the findings, the modified S_1 procedures were robust with the exception of Q_n which generated conservative Type I error rates (below 0.025 value). Besides, S_1 statistic with T_n performed the best among the procedures under skewed distributions.

Cui, He and Ng (2003) proposed an alternative procedure for the analysis of principal components by replacing the classical variability measure such as variance with a robust dispersion measure. For comparison purpose with classical principal components, the three chosen robust estimates of scale were trimmed standard deviation with α = 0.1, Q_n and median absolute deviation, MAD_n. The trivariate samples x_i with the size of 100 and 200 were simulated from the Normal model,

$N\{0, \text{diag}(1, 2, 4)\}$ and the Contaminated-Normal model, $0.9N\{0, \text{diag}(1, 2, 4)\} + 0.1N\{0, \text{diag}(25, 2, 4)\}$. Normal model was used for the purpose of evaluating the efficiency of the robust procedures under strict Gaussian models. Meanwhile, the Contaminated-Normal model was considered for checking the value of the robust principal component analysis procedures relative to the classical principal component analysis when a modest number of outliers existed in the sample. Based on the findings, the authors noticed that the robust procedures had greater bias compared to classical principal component analysis yet they performed well in terms of efficiency. However, MAD_n had a very low level of efficiency than Q_n. In addition, robust procedure with Q_n is the most robust among the three robust scale estimators due to its mean squared errors for estimating the principal component.

2.6 Bootstrap Method

Bootstrapping is a statistical procedure which is also known as computer-based procedure that used to estimate the standard error of $\hat{\theta}$ by resampling the data when the parametric assumptions are in doubt or, parametric inference is either impossible or need a very complicated formula. There were some similar resampling procedures available such as jackknife by Quenouille (1949) and permutation methods by Fisher and Pitman in the 1930s before bootstrap procedure was introduced (Chernick, 1999). Conversely, in year 1979, Efron combined the ideas and linked the simple nonparametric bootstrap which was known as resampling the data with replacement, with the earlier accepted statistical tools like jackknife and delta method. The average weight of people in this world is one of the examples. It is very hard to obtain the weight of people in global population. We can only sample a small part of the population by assuming the size of sample is N. Only one average value can be obtained from that single sample. Bootstrap procedure able us to compute the mean of weight by forming a number of sample sets with replacement from the original data set. This process is repeated with a large numbers of times, normally 1,000, 5,000 or 10,000 times.

According to Chernick (1999), bootstrap can be used as an alternative in certain cases although it may not be providing a very good solution. It is difficult to estimate a parameter, conduct a hypothesis testing about the parameter, determine the standard error or a confidence interval for the parameter with a sample size of n if any parametric assumption is not fulfilled. By considering the empirical distribution which refers to the probability distribution that has probability of $1/n$ assigned to every sample value, the idea of bootstrap is to use it as the replacement for unknown distribution of population. Due to the generality of bootstrap, it has been widely used in many areas than just for estimating the standard errors and confidence intervals.

2.7 Type I Error

Since sample statistics are used to calculate from random data in order to make conclusions about the parameters of populations during the process of hypothesis testing, therefore it is possible that a wrong conclusion would be made with respects to the null hypothesis. Type I error is one of the errors which can occur in testing hypothesis. Type I error is made when the true null hypothesis is rejected. The probability of Type I error is known as alpha (α) or actual level of significance which refers to the area under the curve of the rejection region that beyond the critical value(s). The common values of nominal significance level are 0.05, 0.01 and 0.10 which often decides before the study begins.

Kazempour (1995) studied on the impact of stratification imbalance on the probability of Type I error. The evaluation was carried out in a clinical setting which different response rates of the treatments might be obtained among the strata. The results showed that the dispersion of response rates in a heterogeneous population would affect the Type I error rate when stratification was ignored. The effect on Type I error could be large. Hence, it should be evaluated and addressed. If the response rates in different strata were far apart from each other, the statistical procedures would become more conservative. This applied to no stratification imbalance too.

Brunner and Austin (2007) investigated on the inflation of Type I error in multiple regression when two correlated independent variables were measured with error. Besides, an attempt was made to test one of the independent variables while controlling for the other one by usual regression procedures and measurement error was ignored. The finding proved that Type I error was drastically inflated by ignoring the measurement error in the independent variables of a regression. This outcome could be applied to several types of regression and measurement error due to the failure in making a distinction between the true independent variables. However, the authors claimed that Type I error was not always inflated when measurement error was ignored. It depended on the relationships of independent variables and measurement errors.

Lix *et al*. (1996) mentioned a robust procedure would maintain the actual Type I error rate to be closer to the nominal significance level, α and also maintain the actual statistical power at the same time. Bradley (1978) proposed that if the Type I error rate of a procedure falls in the interval of $0.5\alpha \le \hat{\alpha} \le 1.5\alpha$, the procedure fulfills the criterion of robustness and can be considered robust. For example, when the nominal significance level is set as $\alpha = 0.05$, the procedure is robust if its empirical Type I error rate falls in the range of 0.025 and 0.075. However, according to Guo and Luh (2000), a procedure is robust when the Type I error rate is not more than 0.075 level at $\alpha = 0.05$.

As explained in the previous sections under two-group and more than two-group cases, Type I error rate is affected when the assumptions of normality and variance homogeneity are violated for balanced and unbalanced designs. In our effort to search for a better procedure to overcome the aforementioned problems, in this study we adopt the S_1 statistic that was proposed by Babu *et al*. (1999) for skewed distributions. Instead of using median as the location parameter, we use Hodges-Lehmann estimator in place of the median. We also consider using several robust scale estimators to replace the default scale estimator of S_1 statistic. The detail on S_1 with Hodges-Lehmann will be discussed in the next chapter.

CHAPTER THREE
RESEARCH METHODOLOGY

3.1 Introduction

As mentioned in the previous chapters, S_1 statistic is a statistical procedure that deals with the violation of normality without having to trim or transform the original data. However, the robustness and control of Type I error of this procedure could be further improved especially for skewed distributions. Therefore, the goal of this study is to propose an improved S_1 statistic. The modification on this statistic was done by replacing the default location estimator (median) with Hodges-Lehmann whereas the default scale estimator was substituted by several robust scale estimators.

Most of the real data are non-normal in general (Reed, 1998). Besides, it is a challenge to perform data transformation or data trimming as it might cause the losing of important information of the data. Hence, these wonders can be eliminated by using Hodges-Lehmann as the location parameter because it estimates the "pseudo-median" which is related to population median closely (Boos, 1982). Furthermore, it also can deal with the original data when there is a violation of normality. According to Geyer (2006), the breakdown point of Hodges-Lehmann is 0.3. Although the breakdown point of Hodges-Lehmann is lower than the breakdown point of median, this does not mean median dominates in terms of efficiency for all skewed distributions for example the distributions with relatively light tails (Wilcox, 2012). Unlike median which removes 50% of the largest and 50% of the smallest observations from the data, Hodges-Lehmann takes all the observations from the data into consideration by calculating the midpoints of every two observations before obtaining the median among the computed midpoints.

Conditions such as non-normality and variance heterogeneity are the main concern in this study. Therefore, skewness and kurtosis of the distribution are controlled by using *g* and *h* distribution. Besides, the research design contained two and four groups of data with equal and unequal sample sizes and variances. In addition, the pairing of sample sizes and variances (positive and negative pairing) were tested as well.

3.2 Procedure Employed

For achieving the goal of this study, the location estimator of S_1 statistic, median was replaced by Hodges-Lehmann while for the scale estimators, variance of Hodges-Lehmann, MAD_n, T_n, S_n and Q_n were chosen as the substitution for the default scale estimator, $\hat{\omega}$. The modification of S_1 statistic generates five procedures as shown in Figure 3.1 below.

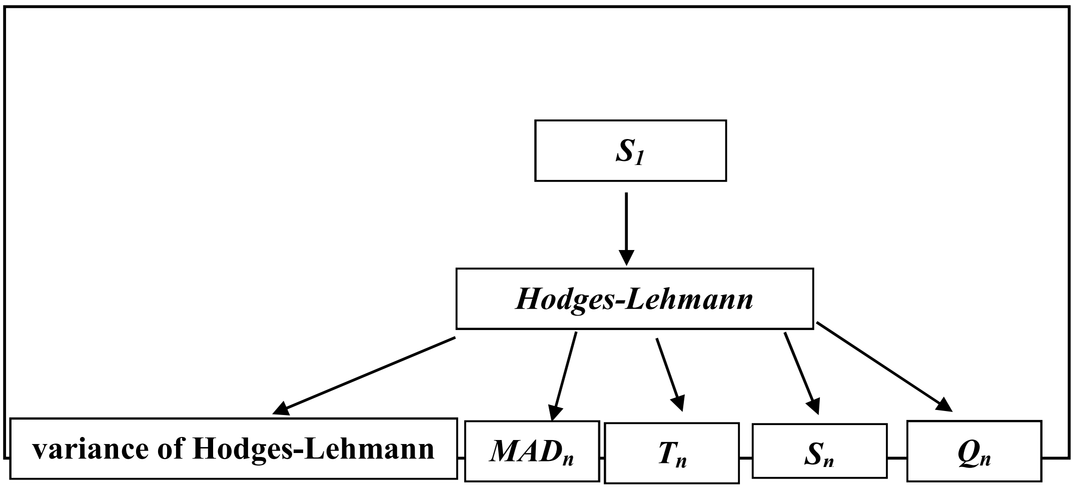

Figure 3.1: Statistical test with the corresponding scale estimators

3.2.1 S_1 using Hodges-Lehmann with its variance

Let $Y_{ij} = (Y_{1j}, Y_{2j}, ..., Y_{nj})$ be a sample from an unknown distribution F_j, j = 1, 2, .., J, where J is the number of groups.

$$S_1 = \sum_{1 \le i < j \le J} |s_{ij}| \tag{3.1}$$

where

$$s_{ij} = \frac{(HL_i - HL_j)}{\sqrt{(\hat{\omega}_i + \hat{\omega}_j)}}; \tag{3.2}$$

HL_i = the Hodges-Lehmann of group i;

HL_j = the Hodges-Lehmann of group j;

$$\hat{\omega}_i = \frac{\omega_i}{n_i}; \hat{\omega}_j = \frac{\omega_j}{n_j}; \tag{3.3}$$

n_i = number of observations for group i;

n_j = number of observations for group j;

$$\omega_i = \left(\frac{1}{n_i} \sum_{i=1}^{I} \sum_{j=1}^{n_i} |Y_{ij} - HL_i| \right)^2 \tag{3.4}$$

$$\omega_j = \left(\frac{1}{n_j} \sum_{j=1}^{J} \sum_{i=1}^{n_j} |Y_{ij} - HL_j| \right)^2 \tag{3.5}$$

3.2.2 S_1 using Hodges-Lehmann with MAD_n

$$S_1 = \sum_{1 \le i < j \le J} |s_{ij}| \tag{3.6}$$

where

$$s_{ij} = \frac{(HL_i - HL_j)}{\sqrt{(MAD_{ni} + MAD_{nj})}}; \tag{3.7}$$

$$MAD_n = b\ med_i \left| Y_{ij} - HL_j \right| \tag{3.8}$$

3.2.3 S_1 using Hodges-Lehmann with T_n

$$S_1 = \sum_{1 \le i < j \le J} \left| s_{ij} \right| \tag{3.9}$$

where

$$s_{ij} = \frac{(HL_i - HL_j)}{\sqrt{(T_{ni} + T_{nj})}}; \tag{3.10}$$

$$T_n = 1.3800 \frac{1}{h} \sum_{m=1}^{h} \left\{ \underset{k \neq i}{med} \left| Y_{ij} - Y_{kj} \right| \right\}_{(m)} \tag{3.11}$$

3.2.4 S_1 using Hodges-Lehmann with S_n

$$S_1 = \sum_{1 \le i < j \le J} \left| s_{ij} \right| \tag{3.12}$$

where

$$s_{ij} = \frac{(HL_i - HL_j)}{\sqrt{(S_{ni} + S_{nj})}}; \tag{3.13}$$

$$S_n = c\ med_i \left\{ med_j \left| Y_{ij} - Y_{kj} \right| \right\} \tag{3.14}$$

3.2.5 S_1 using Hodges-Lehmann with Q_n

$$S_1 = \sum_{1 \le i < j \le J} \left| s_{ij} \right| \tag{3.15}$$

where

$$s_{ij} = \frac{(HL_i - HL_j)}{\sqrt{(Q_{ni} + Q_{nj})}}; \quad (3.16)$$

$$Q_n = d\left\{\left|y_{ij} - y_{kj}\right|; i < k\right\}_{(m)} \quad (3.17)$$

with d as a constant factor,

$$m = \binom{h}{2} \approx \binom{n}{2} \Big/ 4 \text{ and } h = [n/2] + 1 \quad (3.18)$$

3.3 Variables Manipulated

Manipulating variables by creating various conditions are necessary to check on the strengths and weaknesses of the proposed procedures. Further explanations for each condition will be discussed in the following subsections.

3.3.1 Number of Groups

For the two groups case, J which denotes the number of groups is represented by J = 2. Meanwhile, J = 4 represents the more than two groups case since the traditional ANOVA F test was found to perform well under this case (Wilcox, 1994).

3.3.2 Balanced and Unbalanced Sample Sizes

When there are violation on the assumptions of normality and equal variances, inflation of Type I error rates usually occur for parametric procedures. This applies to equal and unequal sample sizes as well (Snedecor and Cochran, 1980). Under equal sample sizes, the effect of non-normality on the Type I error rates of ANOVA F test showed no difference between groups with equal group variances compared to equal sample sizes with unequal group variances. Regardless of the degree of non-normality, the error rates remained close to nominal significance level under equal group variances and inflated across the non-normal distributions under unequal group variances. The same pattern was also found under unequal sample sizes. (Lix *et al*., 1996). Therefore, to prove that the proposed procedures can be used to solve the issue, balanced and unbalanced sample sizes were assigned in this study for two and four groups case.

The total sample sizes for two groups (J = 2) was set as 40. For balanced sample sizes, each group was assigned 20 observations ($n_1 = 20$ and $n_2 = 20$). For unbalanced sample sizes, the first group was set as $n_1 = 15$ and the second group was $n_2 = 25$ (refer to Table 3.2).

Total sample sizes for four groups (J = 4) was twice the total of two groups (J = 2) that was 80 for balanced and unbalanced sample sizes. For balanced sample sizes, each group was pegged with 20 observations per group ($n_1 = n_2 = n_3 = n_4 = 20$). For unbalanced sample sizes, each group was assigned with different number of observations such that $n_1 = 10$, $n_2 = 15$, $n_3 = 25$ and $n_4 = 30$ (refer to Table 3.2).

3.3.3 Types of Distributions

Kang and Harring (2012) investigated the impact of non-normality, effect size and sample size on two groups case with equal sample sizes (n = 8, 15, 30, 60 and 120) for parametric, nonparametric and robust procedures. The results obtained shows that different distributional shapes could affect the control of Type I error of each procedure. Similar effect occurs on more than two groups case as well. Khan and

Rayner (2003) studied on the robustness to non-normality of ANOVA and Kruskal-Wallis for three groups location problem using several types of distributions. Rather than its skewness, both procedures were much affected by the kurtosis of the error distribution. Therefore, to examine the effect of distributional shapes on Type I error, three distributions with different levels of skewness and kurtosis were chosen for this study. A normal distribution represents a distribution with zero skewness. For moderate skewness or mild departure from normality, chi-square distribution with 3 degree of freedom was used. For extreme departure from normality, a skewed heavy tail distribution represented by g = 0.5 and h = 0.5 distribution was used. For this distribution, the skewness is controlled by parameter g while the kurtosis is controlled by parameter h.

3.3.4 Variance Heterogeneity

Variance homogeneity is a crucial assumption when testing for equality of location measures using parametric procedures (Kulinskaya, Staudte and Gao, 2003). The existence of variance heterogeneity across groups could make the statistical procedures for treatment effects unreliable. In this study, variances with different ratios was assigned to two and four groups for equal and unequal sample sizes to investigate the effect of variance heterogeneity on Type I error. For equal sample sizes, variance with 1:1 ratio was allocated to two groups while variance with 1:1:1:1 ratio was allocated to four groups. For unequal sample sizes, variance with 1:36 ratio was assigned to two groups whereas variance with 1:1:1:36 ratio was assigned to four groups (refer to Table 3.3). According to Keselman, Wilcox, Othman & Fradette (2002) and Syed Yahaya (2005), it is reasonable to use the ratio of 1:36 in investigating the procedures' performance under extreme conditions although the ratio may seem large because if a procedure can perform well under such an extreme condition, it can work under most conditions of heterogeneity too.

3.3.5 Nature of Pairings

Nature of pairings are categorised into positive and negative pairings which can only be formed when there are unequal sample sizes that pairs with unequal group variances. Positive pairing is referred to the pairing of the smallest group size with the least group variance and the largest group size is paired with the greatest group variance. For negative pairing, the concept is completely the opposite of positive pairing. The smallest number of group observation is matched with the greatest group variance whereas the largest number of group observation is matched with the lowest group variance (refer to Table 3.4). Generally, positive pairing procedures conservative results and negative pairing produces liberal results (Teh, Md Yusof, Yaacob and Othman, 2010).

3.4 Design Specification

Number of groups, levels of skewness, equal and unequal sample sizes, variance heterogeneity and nature of pairings are the variables that have been manipulated in this study to create various conditions for testing on the strengths and weaknesses of the proposed statistical procedures in controlling Type I error. Basically, the conditions that were created by the manipulation of the variables were further classified into perfect, mild departure and extreme departure from the assumptions. Perfect condition is referred to the condition with equal group sizes, equal group variances and normal distribution. Mild departure focused on equal and unequal group sizes and group variances. The distributional shapes for mild departure are divided into skewed distribution with equal sample sizes and equal group variances and normal distribution with unequal sample sizes and unequal group variances. Extreme departure condition will only concentrate on unequal group sizes and group variances with skewed distributional shape (refer to Table 3.1).

Table 3.1

Conditions of Departure

Conditions	**Group Sizes**	**Group Variances**	**Distributional Shape**
Perfect	Equal	Equal	Normal
Mild Departure	Equal	Equal	Skewed
	Unequal	Unequal	Normal
Extreme Departure	Unequal	Unequal	Skewed

Table 3.2

Sample Sizes

	***J* = 2**		***J* = 4**			
	n_1	n_2	n_1	n_2	n_3	n_4
Equal	20	20	20	20	20	20
Unequal	15	25	10	15	25	30

Table 3.3

Group Variances

	J = 2		*J* = 4			
	n_1	n_2	n_1	n_2	n_3	n_4
Equal	1	1	1	1	1	1
Unequal	1	36	1	1	1	36

Table 3.4

Nature of Pairings

	J = 2		*J* = 4			
	$n_1 = 15$	$n_2 = 25$	$n_1 = 10$	$n_2 = 15$	$n_3 = 25$	$n_4 = 30$
Positive	1	36	1	1	1	36
Negative	36	1	36	1	1	1

3.5 Data Generation

In this study, SAS/IML is the statistical software that used to generate data for different distributional shapes and conditions. The data generation steps for distributional shapes are elaborated as below:

(a) Normal distribution - The mean was set as 0 and standard deviation was set as 1 for the purpose of standardising the normal distribution using SAS generator RANNOR (SAS Institute, 1999).

(b) Chi-square (3 *df*) distribution

i. Produce three standard normal variates Z_{ij} using (a).

ii. Square each Z_{ij}.

iii. Sum up all three squares of Z_{ij}.

$$\chi^2 = \sum_{i=1;\, j=1}^{3} \left(Z_{ij}\right)^2 \tag{3.19}$$

(c) g = 0.5 and h = 0.5 distribution

i. Generate a standard normal variate, Z_{ij} using (a).

ii. Use the equation below to convert the standard normal variates to random variables.

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

Parameter of g is used to control the skewness of the distribution whereas parameter of h is used to control the distribution's kurtosis. With reference to the g-and-h distributions, the skewness increased when g increased and the distribution became heavier when h increased. Therefore, g-and-h = 0.5 represents skewed heavy tail distribution, which can be considered as the extreme departure from normality.

Basically, the values of location parameters are not equal to zero under skewed distributions. Hence, the observations, Y_{ij} from each simulated skewed distributions were centered by subtracting the population location parameter, θ from the observations in order to ensure the null hypothesis, H_0 remains true.

$$X_{ij} = Y_{ij} - \theta \qquad (3.21)$$

According to Othman, *et al*. (2004) and Wilcox and Keselman (2003), the values of θ are determined by calculating $\hat{\theta}$ with one million observations that were simulated from that particular distribution. Precisely, the Hodges-Lehmann of each population distribution has to be subtracted from Y_{ij} in this study. Table 3.5 below represents the central tendency measure (Hodges-Lehman values) with respect to distributions based on one million observations.

Table 3.5

Central Tendency Measure with respect to Distributions

	Normal	**Chi-square**	***g* = 0.5 and *h* = 0.5**
Hodges-Lehmann	0	2.674	0.101

During the generation of data using SAS/IML, the groups central tendency measures was set at $(\theta_1, \theta_2) = (0{,}0)$ for two groups ($J = 2$) and $(\theta_1, \theta_2, \theta_3, \theta_4) = (0{,}0{,}0{,}0)$ for four groups ($J = 4$) for the analysis of Type I error.

3.6 Bootstrap Method

Bootstrap method was used in this study for conducting the hypothesis testing since the sampling distribution for the proposed statistics is unknown. During the process of bootstrap resampling, artificial samples were drawn from the sample itself with replacement. Furthermore, the sampling distribution of the test statistic under the null hypothesis could be estimated by the bootstrap. Other than that, the bootstrap was able to locate the rejection region of the test empirically. Monte Carlo procedure was used for the purpose of obtaining Type I error. All data sets of each design (refer to Table 3.2 to Table 3.4) were simulated for 5000 times at 5% level of significance ($\alpha = 0.05$). Below are the steps for testing hypothesis using bootstrap method.

i. Use the original data that are produced by SAS/IML to calculate S_1.

ii. Bootstrap the original data in order to get the sample bootstrap data.

iii. Centralise the data.

iv. Use bootstrap data to calculate S_1 and denote it as S_1^*.

v. Repeat the steps from (ii) to (iv) for $B = 599$.

vi. Calculate p-value using (number of $S_1^* > S_1$) / B.

The formula to calculate Type I error rate is the number of p-values that is less than 0.05 divided by the number of simulation. For example, if the data sets are simulated for 5000 times, Type I error rate is computed as (number of p-values < 0.05) / 5000.

CHAPTER FOUR
RESULTS OF THE ANALYSIS

4.1 Introduction

The structure of this chapter focuses on the study of the proposed test statistic, S_1 statistic by replacing the default location parameter, median with Hodges-Lehmann. Other than the variance of Hodges-Lehmann, four highly robust scale estimators, MAD_n, T_n, S_n and Q_n were also chosen to replace the default scale estimator, $\hat{\omega}$ in S_1 statistic. As mentioned in Section 3.2, five procedures were developed in this study. In order to accomplish the objectives that are mentioned in Section 1.3, their robustness in terms of Type I error rates will be compared. For the purpose of showing their strengths and weaknesses, several conditions were created for each of these procedures such as the levels of skewness, equal and unequal group variances, equal and unequal sample sizes, and nature of pairings. The nature of pairings is occurred for unbalanced design only which is referred to unequal group sizes and unequal group variances. All conditions were arranged under balanced and unbalanced group designs under two groups (J = 2) and four groups (J = 4) for each procedure. The empirical rate of Type I error obtained are presented in this chapter.

Normal, chi-square with three degrees of freedom and g-and-h distribution were used to represent the different levels of skewness which are referred to zero, mild and extreme skewness. For the ease of comparison, the results of Type I error rates are arranged in table form according to each procedure and distributional shape. Type of distribution is placed as the first column followed by each procedure in the next five columns for balanced group design of J = 2 and J = 4. Meanwhile, for unbalanced design of J = 2 and J = 4, nature of pairing which is positive or negative pairing is presented in the second column followed by the five proposed procedures. Pairings only exist for unbalanced group design.

In order to test the robustness of each proposed procedure under every condition for the 5% of significance level $(\alpha = 0.05)$ that used in this study, Guo and Luh's criterion of robustness was employed as the benchmark for measuring the robustness in terms of Type I error. According to Guo and Luh (2000), a procedure is considered robust if its empirical Type I error rate does not greater than 0.075 at $\alpha = 0.05$. Besides, a procedure that produces Type I error rate which is closest to nominal significance level can be called as the best procedure in controlling Type I error.

4.2 S_1 Procedures

The central tendency measure of S_1 statistic, median was substituted by Hodges-Lehmann in this study. Five scale estimators mentioned in previous section were each integrated in the S_1 statistic forming five proposed procedures namely S_1 with the variance of Hodges-Lehmann, $S_1(HL)$; S_1 with MAD_n, $S_1(MAD_n)$; S_1 with T_n, $S_1(T_n)$; S_1 with S_n, $S_1(S_n)$ and S_1 with Q_n, $S_1(Q_n)$, The goal of the proposed S_1 statistic is to test the equality of Hodges-Lehmann such that:

$$H_0 : HL_1 = HL_2 = \ldots = HL_J$$

where HL_j is the population Hodges-Lehmann for group j such as j = 1, 2, …, J. Percentile bootstrap method was used to generate Type I error rates for all the five proposed S_1 procedures. The analysis of Type I error is explained in section 4.2.1 and

4.2.2 based on two groups (J = 2) and four groups (J = 4) cases under balanced and unbalanced group designs.

4.2.1 Type I Error for *J* = 2

The null hypothesis for two groups (J = 2) is given as:

$$H_0 : HL_1 = HL_2$$

The Type I error rates obtained for each modified S_1 statistic are arranged in Table 4.1 (section 4.2.1.1) and 4.2 (section 4.2.1.2) for balanced and unbalanced group designs.

4.2.1.1 Balanced Design (*J* = 2)

Balanced design referred to the groups having equal number of observations and homogeneous group variances. The results obtained are presented in Table 4.1 below.

Table 4.1

Type I Error Rates for J = 2 (Balanced Design)

Distribution	S_1 (*HL*)	S_1 (MAD_n)	S_1 (T_n)	S_1 (S_n)	S_1 (Q_n)
Normal	0.0304	0.0360	0.0350	0.0356	0.0266
Chi-square	0.0262	0.0250	0.0272	0.0282	0.0220
g = 0.5, h = 0.5	0.0074	0.0212	0.0186	0.0190	0.0124

The Type I error rates of each proposed procedure across all three distributions fulfill the Guo and Luh's robust criterion as all the Type I error rates fall below 0.075 level.

Based on the results obtained, $S_1(MAD_n)$ has better control of Type I error for normal (0.0360) and *g*-and-*h* = 0.5 distribution (0.0212) because its empirical Type I error rates are the closest to the nominal level of 0.05. In contrast, $S_1(S_n)$ performs the best for chi-square (0.0282). Meanwhile, $S_1(Q_n)$ shows the lowest (furthermost from the nominal level of 0.05) Type I error for normal (0.0266) and chi-square (0.0220). As for the *g*-and-*h* = 0.5, $S_1(HL)$ produce the lowest Type I error (0.0074), and this value seems to be the lowest among all conditions. The range of values across the table spans from 0.0074 to 0.0360. In general, $S_1(MAD_n)$ shows better performance (nearest) to the nominal level) than the other procedures while $S_1(Q_n)$ produces very low Type I error rates (further from the nominal level). Anyhow, the results for all conditions are robust.

4.2.1.2 Unbalanced Design ($J = 2$)

Table 4.2 below presents the Type I error rates obtained for $S_1(HL)$, $S_1(MAD_n)$, $S_1(T_n)$, $S_1(S_n)$ and $S_1(Q_n)$ under three distributions according to the positive and negative pairings.

Table 4.2

Type I Error Rates for J = 2 (Unbalanced Design)

Distribution	**Nature Of Pairing**	S_1 **(*HL*)**	S_1 **(MAD_n)**	S_1 **(T_n)**	S_1 **(S_n)**	S_1 **(Q_n)**
Normal	Positive	0.0288	0.0382	0.0390	0.0404	0.0302
	Negative	0.0164	0.0254	0.0276	0.0202	0.0140
Chi-square	Positive	0.0234	0.0442	0.0460	0.0450	0.0360

	Negative	0.0174	0.0348	0.0402	0.0292	0.0258

Table 4.2 *continued*

g = 0.5, h = 0.5	Positive	0.0104	0.0296	0.0304	0.0282	0.0192
	Negative	0.0058	0.0200	0.0194	0.0132	0.0088

Refer to Table 4.2 above, it can be clearly seen that all procedures are robust as none of the Type I error rates fall above 0.075 level according to Guo and Luh's criterion of robustness. For positive pairing, $S_1(S_n)$ shows better control of Type I error for under normal distribution (0.0404) whereas $S_1(T_n)$ performs the best for chi-square (0.0460) and g-and-h = 0.5 (0.0304) because their Type I error rates are the closest to nominal level of 0.05. In the meantime, for negative pairing, $S_1(T_n)$ shows the best performance for normal (0.0276) and chi-square distributions (0.0402) while for g-and-h = 0.5, $S_1(MAD_n)$ has the greatest and closest error rate to nominal level (0.0200). On the other hand, $S_1(Q_n)$ produces the lowest empirical Type I error rate for negative pairing under normal distribution (0.0140) which is the furthest from nominal level. Regardless of the nature of pairing and distributional shape, $S_1(HL)$ is ranked last for all conditions except for negative pairing under normal distribution. The empirical Type I error rates in Table 4.2 are ranging from 0.0058 to 0.0460. We observe that this range of value is wider than the range of value under balanced design. In general, $S_1(T_n)$ shows the best performance in controlling Type I error (near to nominal level). Meanwhile, $S_1(HL)$ generates very low Type I error rates (further from nominal level). However, all procedures are robust across all conditions.

4.2.2 Type I Error for J = 4

For four groups case, the null hypothesis is given as:

$$H_0 : HL_1 = HL_2 = HL_3 = HL_4$$

The Type I error rates are presented in Table 4.3 (section 4.2.2.1) and 4.4 (section 4.2.2.2) below using the same arrangement as in section 4.2.1.1 and 4.2.1.2 for $J = 2$.

4.2.2.1 Balanced Design (J = 4)

Similar to the balanced design for J = 2, the tests were conducted based on the groups that have equal sample sizes and group variances. The results of each distributional shape are shown in Table 4.3 below.

Table 4.3

Type I Error Rates for J = 4 (Balanced Design)

Distribution	S_1 **(*HL*)**	S_1 **(*MAD$_n$*)**	S_1 **(T_n)**	S_1 **(S_n)**	S_1 **(Q_n)**
Normal	0.0220	0.0220	0.0206	0.0234	0.0138
Chi-square	0.0128	0.0118	0.0120	0.0102	0.0070
g = 0.5, h = 0.5	0.0022	0.0096	0.0068	0.0064	0.0034

The Type I error rates present in Table 4.3 shows that all the proposed procedures, $S_1(HL)$, $S_1(MAD_n)$, $S_1(T_n)$, $S_1(S_n)$ and $S_1(Q_n)$ fulfill Guo and Luh's criterion of robustness by obtaining Type I error rates that fall below 0.075 level. For normal distribution, the procedure shows better performance is $S_1(S_n)$ (0.0234) by having the closest Type I error rate to nominal level. Meanwhile, $S_1(HL)$ has the best control of Type I error among the procedures for chi-square with (0.0128). For g-and-h = 0.5, $S_1(MAD_n)$ (0.0096) performs the best (closest to nominal level) even though the value is pretty small as all the Type I error rates of the procedures under this distribution deflate to below 0.01. In contrast, $S_1(Q_n)$ produced the lowest error rates for normal (0.0138) and chi-square distributions (0.0070) (furthermost from nominal level). This outcome is the same as in balanced design under two groups case. The result also clearly shown that the empirical Type I error rate of $S_1(HL)$ is again the lowest for g-and-h = 0.5 (0.0022) and also among all conditions. The range of Type I error rates in Table 4.3 spans from 0.0022 to 0.0234. In general, for J = 4 balanced design, the best procedure in controlling Type I error is $S_1(MAD_n)$ (close to nominal level) while $S_1(Q_n)$ produces very low Type I error rates (further from nominal level). Nevertheless, all procedures are robust.

4.2.2.2 Unbalanced Design (J = 4)

For the case of four groups under unbalanced design, the Type I error rates for $S_1(HL)$, $S_1(MAD_n)$, $S_1(T_n)$, $S_1(S_n)$ and $S_1(Q_n)$ are displayed in Table 4.4.

Table 4.4

Type I Error Rates for J = 4 (Unbalanced Design)

Distribution	**Nature of Pairing**	S_1 **(*HL*)**	S_1 **(*MAD*$_n$)**	S_1 **(*T*$_n$)**	S_1 **(*S*$_n$)**	S_1 **(*Q*$_n$)**

Normal	Positive	0.0230	0.0252	0.0234	0.0194	0.0162
	Negative	0.0178	0.0196	0.0180	0.0160	0.0112
Chi-square	Positive	0.0148	0.0226	0.0260	0.0230	0.0192
	Negative	0.0158	0.0272	0.0256	0.0260	0.0180

Table 4.4 *continued*

$g = 0.5$, $h = 0.5$	Positive	0.0028	0.0150	0.0134	0.0120	0.0066
	Negative	0.0032	0.0130	0.0108	0.0098	0.0042

According to the results above, $S_1(HL)$, $S_1(MAD_n)$, $S_1(T_n)$, $S_1(S_n)$ and $S_1(Q_n)$ are considered robust due to the criterion of robustness by Guo and Luh (2000). With the exception of positive pairing under chi-square (3 *df*) distribution, $S_1(MAD_n)$ shows the best performance regardless of the nature of pairing across all three distributions due to its Type I error rates closest to the 0.05 nominal level. Meanwhile, $S_1(T_n)$ produces the highest empirical Type I error rate for positive pairing under chi-square (3 *df*) distribution (0.0260). In contrast, $S_1(Q_n)$ is ranked last for positive (0.0162) and negative (0.0112) pairings under normal distribution. For the conditions of mild (chi-square) and extreme skewness (*g*-and-*h* = 0.5), $S_1(HL)$ again performs the worst with regards of the nature of pairing. Furthermore, the values for *g*-and-*h* = 0.5 are also the smallest among all conditions. The range of the empirical Type I error rates obtained for four groups case (unbalanced design) spans from 0.0028 to 0.0272. In general, it seems that $S_1(MAD_n)$ is the best procedure in controlling Type I error than other procedures while $S_1(HL)$ has the lowest Type I error rates. Yet, all procedures are considered robust.

Due to the consistency of generating the closest empirical Type I error rates to the nominal level of 0.05 for balanced and unbalanced designs under two and four groups cases, $S_1(MAD_n)$ is the best procedure in controlling Type I error when the levels of skewness are zero, mild and extreme. In the meantime, $S_1(Q_n)$ shows the worst performance with its lowest Type I error rates that are furthermost from the nominal level for balanced design under two and four groups cases. For unbalanced design, the worst procedure in controlling Type I error is $S_1(HL)$ under two and four groups cases. Nonetheless, all the proposed procedures, $S_1(HL)$, $S_1(MAD_n)$, $S_1(T_n)$, $S_1(S_n)$ and $S_1(Q_n)$ are robust among all conditions for both group designs and cases.

4.3 S_1 Statistic versus Parametric and Nonparametric Procedures

In order to determine the suitability of $S_1(HL)$, $S_1(MAD_n)$, $S_1(T_n)$, $S_1(S_n)$ and $S_1(Q_n)$ to be the alternative procedures in testing the equality of location measures for zero, mild and extreme skewness, the comparison based on Type I error rates between all the five proposed procedures with parametric and nonparametric procedures is necessary. This comparison is conducted according to the group case ($J = 2$ or $J = 4$) and group design (balanced and unbalanced) for each distributional shape. For $J = 2$, the parametric procedure is *t*-test, while Mann-Whitney is the choice of nonparametric procedure for this study. For $J = 4$, the chosen parametric and nonparametric procedures are ANOVA and Kruskal-Wallis respectively. At the same time, each proposed procedure is also compared with the original S_1 statistic, $S_1(\hat{\omega})$. For the ease of comparison among procedures in terms of empirical Type I error rates, there is a slight change on the structure of the results table. For balanced group design, the procedures are displayed in the first column followed by the columns of distributions. As for the unbalanced design, nature of pairing is displayed in the second column followed by the distribution columns.

4.3.1 Type I Error *J* = 2 (Balanced Design)

The empirical Type I error rates for $S_1(HL)$, $S_1(MAD_n)$, $S_1(T_n)$, $S_1(S_n)$, $S_1(Q_n)$, $S_1(\hat{\omega})$, *t*-test and Mann-Whitney are displayed in Table 4.5.

Table 4.5

Type I Error Rates for J = 2 (Balanced Design)

Procedure	Distribution		
	Normal	**Chi-Square**	**g = 0.5, h = 0.5**
S_1 (HL)	0.0304	0.0262	0.0074
S_1 (MAD_n)	0.0360	0.0250	0.0212
S_1 (T_n)	0.0350	0.0272	0.0186
S_1 (S_n)	0.0356	0.0282	0.0190
S_1 (Q_n)	0.0266	0.0220	0.0124
S_1 ($\hat{\omega}$)	0.0364	0.0342	0.0148
t-test	0.0528	0.0500	0.0288
Mann-Whitney	0.0526	0.0566	0.0526

Based on the values presented in Table 4.5, the five proposed procedures, $S_1(\hat{\omega})$, *t*-test and Mann-Whitney are considered robust because all values fall below 0.075

level for normal, chi-square (3 *df*) and *g*-and-*h* = 0.5 distributions. Mann-Whitney performs the best in controlling Type I error under normal (0.0526) and *g*-and-*h* = 0.5 distributions (0.0526) whereas *t*-test provides the best result under chi-square (0.0500) as their Type I error rates are the closest to nominal level of 0.05. Meanwhile, the procedures that generate the lowest empirical Type I error rates are $S_1(Q_n)$ for normal and chi-square distribution, and $S_1(HL)$ for *g*-and-*h* = 0.5 (furthermost from the nominal level). The range of values in Table 4.5 spans from 0.0074 to 0.0566.

In comparing the parametric (*t*-test), nonparametric (Mann-Whitney) and the original S_1 procedure, $S_1(\hat{\omega})$ against the five proposed procedures, $S_1(MAD_n)$ (0.0212), $S_1(S_n)$ (0.0190) and $S_1(T_n)$ (0.0186) produce better Type I error rates than $S_1(\hat{\omega})$ (0.0148) under *g*-and-*h* = 0.5 distribution. However, these three proposed procedures are still ranked behind Mann-Whitney and *t*-test. For normal and chi-square distributions, none of the proposed procedures show better control of Type I error than *t*-test, Mann-Whitney and $S_1(\hat{\omega})$.

In general, Mann-Whitney shows the best performance than the other procedures in terms of controlling Type I error for two groups case under balanced design (closest to nominal level) while $S_1(Q_n)$ has the lowest Type I error rates. However, all procedures are considered robust. In addition, the goal of this study has been achieved as the three proposed procedures, $S_1(MAD_n)$, $S_1(S_n)$ and $S_1(T_n)$ improve the original S_1 statistic for extremely skewed distribution.

4.3.2 Type I Error $J = 2$ (Unbalanced Design)

The Type I error rates for two groups under unbalanced group design have been recorded in Table 4.6 below. Unbalanced design is meant for unequal sample sizes, unequal group variances and nature of pairings.

Table 4.6

Type I Error Rates for J = 2 (Unbalanced Design)

Procedure	**Nature of pairing**	**Distribution**		
		Normal	**Chi-Square**	$g = 0.5, h = 0.5$
S_1 (*HL*)	Positive	0.0288	0.0234	0.0104
	Negative	0.0164	0.0174	0.0058
S_1 (MAD_n)	Positive	0.0382	0.0442	0.0296
	Negative	0.0254	0.0348	0.0200
S_1 (T_n)	Positive	0.0390	0.0460	0.0304
	Negative	0.0276	0.0402	0.0194
S_1 (S_n)	Positive	0.0404	0.0450	0.0282
	Negative	0.0202	0.0292	0.0132
S_1 (Q_n)	Positive	0.0302	0.0360	0.0192
	Negative	0.0140	0.0258	0.0088
S_1 ($\hat{\omega}$)	Positive	0.0448	0.0426	0.0192
	Negative	0.0422	0.0390	0.0156
t-test	Positive	0.0198	0.0238	0.0118
	Negative	0.1268	0.1678	0.1048

Mann-Whitney	Positive	0.0448	0.0666	0.0426
	Negative	0.1086	0.1312	0.0976

According to the Type I error rates that is presented above, all procedures are considered robust as their Type I error rates fall below 0.075 level with regards to the nature of pairing and distributional shape, except for *t*-test and Mann-Whitney under negative pairing across all three distributions. For positive pairing, Mann-Whitney emerged as the best performer for normal (0.0448) and *g*-and-*h* = 0.5 distributions (0.0426) as these values are the closest to nominal level of 0.05. Besides, $S_1(\hat{\omega})$ also performed exceptionally well for normal distribution with the same Type I error rate as Mann-Whitney. On the other hand, $S_1(T_n)$ has better control of Type I error for chi-square (0.0460). For negative pairing, $S_1(\hat{\omega})$ produces the closest Type I error rate to nominal level for normal distribution (0.0422) whereas $S_1(T_n)$ shows the best performance for chi-square (0.0402). For *g*-and-*h* = 0.5, $S_1(MAD_n)$ has the best empirical Type I error rate (0.0200) which is the closest to nominal level. Meanwhile, $S_1(HL)$ generates the lowest (furthermost from nominal level) Type I error rates across all conditions except for negative pairing under normal distribution. The procedure that is ranked last for this particular condition is $S_1(Q_n)$. The Type I error rates obtained fall in the range of 0.0058 to 0.1678.

From Table 4.6, we observe that for positive pairings under chi-square and *g*-and-*h* = 0.5 distributions, $S_1(T_n)$ (0.0460; 0.0304), $S_1(S_n)$ (0.0450; 0.0282) and $S_1(MAD_n)$ (0.0442; 0.0296) have better control in terms of Type I error than $S_1(\hat{\omega})$ (0.0426; 0.0192). For negative pairings under chi-square and *g*-and-*h* = 0.5 distributions,

$S_1(T_n)$ (0.0402; 0.0194) generates Type I error rate which is closer to nominal significance level of 0.05 than $S_1(\hat{\omega})$ (0.0390; 0.0156). Besides $S_1(T_n)$, $S_1(MAD_n)$ (0.0200) also performs better than $S_1(\hat{\omega})$ for g-and-h = 0.5 distribution (negative pairing).

Generally, t-test does not perform well in controlling Type I error for positive pairing across the three distributions. When the level of skewness is zero, it produces the lowest rate (0.0198), while ranked the second lowest under chi-square (0.0238) and extremely skewed distribution with g-and-h = 0.5 (0.0118). In contrast, although Mann-Whitney produces the Type I error rate of 0.0666 for positive pairing under chi-square distribution, it does not emerge as the best procedure in controlling Type I error compares to $S_1(T_n)$, $S_1(MAD_n)$ and $S_1(S_n)$ as well as $S_1(\hat{\omega})$.

Overall, with respects to distributional shape, Mann-Whitney emerges as the best procedure for unbalanced design under two groups case for positive pairings and $S_1(MAD_n)$ has the best Type I error rates for negative pairings (nearest to nominal level) while $S_1(HL)$ is the worst procedure in controlling Type I error (further from nominal level). Nevertheless, all procedures can be considered robust regardless of nature of pairing and distributional shape except for parametric and nonparametric procedures for negative pairings across all the three levels of skewness. Besides, the proposed procedures, $S_1(MAD_n)$, $S_1(S_n)$ and $S_1(T_n)$ have achieved the goal of this study by improving the performance of the original S_1 statistic in terms of controlling Type I error for skewed distributions.

4.3.3 Type I Error J = 4 (Balanced Design)

Table 4.7 displays the Type I error rates of $S_1(HL)$, $S_1(MAD_n)$, $S_1(T_n)$, $S_1(S_n)$, $S_1(Q_n)$, $S_1(\hat{\omega})$, ANOVA and Kruskal-Wallis for four groups under balanced (equal group sizes and equal group variances) design.

Table 4.7

Type I Error Rates for J = 4 (Balanced Design)

Procedure	**Distribution**		
	Normal	**Chi-Square**	***g* = 0.5, *h* = 0.5**
$S_1\ (HL)$	0.0220	0.0128	0.0022
$S_1\ (MAD_n)$	0.0220	0.0118	0.0096
$S_1\ (T_n)$	0.0206	0.0120	0.0068
$S_1\ (S_n)$	0.0234	0.0102	0.0064
$S_1\ (Q_n)$	0.0138	0.0070	0.0034
$S_1\ (\hat{\omega})$	0.0230	0.0130	0.0058
ANOVA	0.0510	0.0450	0.0290
Kruskal-Wallis	0.0498	0.0440	0.0498

Based on the results in Table 4.7, all procedures are considered robust according to Guo and Luh's criterion of robustness as each error rate falls below 0.075 level. Kruskal-Wallis shows the best performance for normal (0.0498) and *g*-and-*h* = 0.5 distributions (0.0498) because the values are the closest to nominal level of 0.05.

Meanwhile, ANOVA has the best control of Type I error for chi-square (nearest to nominal level). Those two procedures are ranked either first or second across all conditions. $S_1(Q_n)$ is the worst procedure for normal (0.0138) and chi-square distributions (0.0070) whereas $S_1(HL)$ has the lowest Type I error rate for g-and-h = 0.5 (0.0022) (furthermost from nominal level). The Type I error rates in Table 4.7 fall in the range of 0.0022 to 0.0498.

When the original S_1 procedure, $S_1(\hat{\omega})$ is compared against the five proposed procedures, it can be seen that $S_1(S_n)$ (0.0234) generates nearer Type I error rate to nominal level of 0.05 than $S_1(\hat{\omega})$ (0.0230) for normal distribution. Furthermore, $S_1(MAD_n)$ (0.0096), $S_1(T_n)$ (0.0068) and $S_1(S_n)$ (0.0064) show better performance than $S_1(\hat{\omega})$ (0.0058) when the distribution is extremely skewed. However, $S_1(\hat{\omega})$ produces the closest empirical Type I error rate (0.0130) to the nominal level than the five proposed procedures for chi-square distribution.

In general, Kruskal-Wallis is the best procedure in controlling Type I error (closest to the nominal level) for balanced design under four groups case while $S_1(Q_n)$ procedure produces the lowest Type I error rates (further from nominal level). Like the other designs, all the procedures are robust across all conditions. In addition, $S_1(MAD_n)$, $S_1(T_n)$ and $S_1(S_n)$ again achieve the goal of this study by showing improvement in the S_1 statistic for extremely skewed distribution.

4.3.4 Type I Error J = 4 (Unbalanced Design)

The Type I error rates of $S_1(HL)$, $S_1(MAD_n)$, $S_1(T_n)$, $S_1(S_n)$, $S_1(Q_n)$, $S_1(\hat{\omega})$, ANOVA and Kruskal- Wallis for four groups under unbalanced design (unequal group sizes and unequal group variances) are shown in Table 4.8 below. The pairing of unequal group sizes with unequal group variances yield the positive and negative pairings under the column of nature of pairings.

Table 4.8

Type I Error Rates for J = 4 (Unbalanced Design)

Procedure	**Nature of pairing**	**Distribution**		
		Normal	**Chi-Square**	**g = 0.5, h = 0.5**
S_1 (HL)	Positive	0.0230	0.0148	0.0028
	Negative	0.0178	0.0158	0.0032
S_1 (MAD_n)	Positive	0.0252	0.0226	0.0150
	Negative	0.0196	0.0272	0.0130
S_1 (T_n)	Positive	0.0234	0.0260	0.0134
	Negative	0.0180	0.0256	0.0108
S_1 (S_n)	Positive	0.0194	0.0230	0.0120
	Negative	0.0160	0.0260	0.0098
S_1 (Q_n)	Positive	0.0162	0.0192	0.0066
	Negative	0.0112	0.0180	0.0042
S_1 ($\hat{\omega}$)	Positive	0.0278	0.0246	0.0078
	Negative	0.0302	0.0278	0.0102

ANOVA	Positive	0.0336	0.0526	0.1492
	Negative	0.2850	0.2976	0.3554
Kruskal-Wallis	Positive	0.0448	0.0492	0.0498
	Negative	0.1158	0.1180	0.1022

According to the results in Table 4.8 above, the five proposed procedures and the original S_1 procedure, $S_1(\hat{\omega})$ are robust based on Guo and Luh's criterion of robustness for both positive and negative pairings across all three distributional shapes because their Type I error rates fall below 0.075 level. ANOVA is robust only for positive pairing under normal and chi-square distributions whereas Kruskal-Wallis is robust for positive pairing regardless of the level of skewness.

When the procedures were tested under positive pairing, Kruskal-Wallis shows the best performance in controlling Type I error under normal (0.0448), chi-square (0.0492) and g-and-h = 0.5 (0.0498) distributions as its Type I error rates are the closest to nominal level of 0.05. In contrast, for negative pairing, $S_1(\hat{\omega})$ emerges as the best performer for normal (0.0302) and chi-square distributions (0.0278) while $S_1(MAD_n)$ has the closest Type I error rate to nominal level for g-and-h = 0.5 (0.0130). Meanwhile, $S_1(Q_n)$ produces the most conservative Type I error rates for both nature of pairings under normal distribution. For chi-square and g-and-h = 0.5, $S_1(HL)$ has the lowest empirical Type I error rates for positive and negative pairings (furthermost from nominal level). The range of values across Table 4.8 spans from 0.0028 to 0.3554.

Based on the Type I error rates obtained, $S_1(T_n)$ shows better control of Type I error than the original S_1 procedure, $S_1(\hat{\omega})$ for positive pairing under chi-square when the proposed procedures are compared against $S_1(\hat{\omega})$. Meanwhile, for g-and-h = 0.5, $S_1(MAD_n)$, $S_1(T_n)$ and $S_1(S_n)$ generate greater and closer Type I error rates to the nominal level of 0.05 for positive pairing while $S_1(MAD_n)$ and $S_1(T_n)$ are the best performers for negative pairing compare to $S_1(\hat{\omega})$. On the other hand, ANOVA is ranked second after Kruskal-Wallis for positive pairing under normal (0.0336) and chi-square distributions (0.0526). None of the proposed procedures performs better than ANOVA and Kruskal-Wallis for those two distributional shapes.

Overall, Kruskal-Wallis is the best procedure in controlling Type I error than other procedures for positive pairing whereas $S_1(MAD_n)$ has the best Type I error rates for negative pairings (nearest to nominal level). At the same time, $S_1(HL)$ emerges as the worst procedure due to its very low Type I error rates (further from nominal level). However, all procedures are robust across all conditions with the exception of parametric and nonparametric procedures for negative pairing in all levels of skewness. In addition, $S_1(T_n)$, $S_1(S_n)$ and $S_1(MAD_n)$ show better control of Type I error than the original S_1 procedure, $S_1(\hat{\omega})$ when the level of skewness is extreme.

Refer to the results presented in Table 4.5 to Table 4.8, we observe that the original S_1 and the five proposed procedures are robust according to Guo and Luh (2000) in terms of robustness for balanced and unbalanced designs under two groups and four groups across all three levels of skewness. Parametric procedures, t-test (J = 2) and ANOVA (J = 4) and nonparametric procedures, Mann-Whitney (J = 2) and Kruskal-

Wallis ($J = 4$) in this study are found to be able to control Type I error under positive pairing only for balanced and unbalanced designs.

Meanwhile, among the S_1 procedures, three of the proposed procedures, $S_1(MAD_n)$, $S_1(T_n)$ and $S_1(S_n)$ show an improvement on S_1 statistic in terms of controlling Type I error under skewed distributions for balanced and unbalanced designs under both cases. Furthermore, $S_1(MAD_n)$ provides the best results among these three proposed procedures (closest to nominal level of 0.05). Hence, the goal of this study has been achieved.

4.4 Application on Real Data

Seeking for alternative procedures is always the main concern of most researchers due to the limitations of existing statistical procedures when the assumptions of normality and equal variances are not met. Therefore, in order to ensure the proposed procedures are suitable to be used on data with some violation of assumptions, testing the procedures using real data is necessary.

Stigler (1977) had an enquiry about the efficiency of robust estimators using real data since the advantage of using simulation was clearly known for over 40 years. However, there was no guarantee that the generated pseudo-samples were the actual representative of real data although sampling distributions could be decided through the computer software. Therefore, the author chose eleven robust estimators for comparison purposes and twenty sets of data from 1798 measurements of the earth's mean density, from 1761 determinations of the sun's parallax, and from circa 1880 measurements of the light's speed. According to his finding, the real data showed different behavior from the simulation data which were used in most studies of robustness and this affected the consequent recommendations for the choice of an estimator and also the assessments of the relative performance of estimators.

In this study, the real data, which is listed in Table 4.9 was collected from four lecturers who taught Decision Analysis course at Universiti Utara Malaysia (UUM) seven years ago. The first column of the table represents the group numbers with the quantity of students, whereas, the second column shows the scores of those students obtained in each group.

Table 4.9

Real Data

Group	**Score**
Group 1 (n = 33)	66, 60, 80, 74, 94, 71, 90, 90, 78, 65, 7, 69, 74, 82, 71, 66, 79, 56, 69, 68, 81, 73, 74, 76, 78, 74, 71, 55, 48, 78, 81, 88, 89
Group 2 (n = 19)	69, 69, 57, 65, 86, 57, 71, 71, 70, 74, 65, 67, 67, 90, 73, 85, 56, 74, 66
Group 3 (n = 24)	96, 62, 81, 75, 80, 66, 60, 75, 65, 85, 76, 71, 61, 83, 82, 65, 73, 62, 92, 60, 90, 66, 70, 65
Group 4 (n = 20)	93, 89, 85, 81, 81, 73, 85, 68, 73, 79, 73, 77, 75, 84, 73, 83, 78, 79, 80, 77

Table 4.10

Test of Normality

	Group 1	**Group 2**	**Group 3**	**Group 4**
p-value	0.000	0.130	0.101	0.867

Table 4.11

Test of Homogeneity of Variances

	Levene Statistic	**df1**	**df2**	**Significance**
Score	2.022	3	92	0.116

Table 4.10 presents the result of normality check for each group. At 5% level of significance, the distributions of Group 2, 3 and 4 seem significantly normal. Group 1 is the only group that is not normally distributed with its *p*-value of 0.000 because of the existence of outlier in the data set. In the meantime, the *p*-values in Table 4.11 shows that the assumption of variance homogeneity has been met due to the *p*-value obtained from Levene test is 0.116 which is greater than 0.05.

The real data was tested with the proposed procedures ($S_1(HL)$, $S_1(MAD_n)$, $S_1(T_n)$, $S_1(S_n)$ and $S_1(Q_n)$) as well as the parametric ANOVA, the nonparametric Kruskal-Wallis and the original $S_1(\hat{\omega})$. All the *p*-values are displayed in Table 4.12 below.

Table 4.12

p-value of Real Data (J = 4)

$S_1(HL)$	$S_1(MAD_n)$	$S_1(T_n)$	$S_1(S_n)$	$S_1(Q_n)$	$S_1(\hat{\omega})$	ANOVA	Kruskal-Wallis
0.0167	0.0200	0.0167	0.0267	0.0401	0.0267	0.0790	0.015

In reference to the results displayed in Table 4.11, we observe that Kruskal-Wallis, the original S_1 and the five proposed procedures show the significance in testing the central tendency measures for unequal group sizes and equal group variances at 0.05 significance level for four groups case. ANOVA (p = 0.0790) is the only procedure that fails to show the significance of the test at the 0.05 level. In this case, Kruskal-Wallis provides the most significant result with its lowest p-value (0.015) among all the tested procedures. Tailing closely is $S_1(HL)$ (0.0167) and $S_1(T_n)$ (0.0167). When compared among the S_1 procedures, except for $S_1(Q_n)$, all the other proposed procedures show better or equal significance with the original $S_1(\hat{\omega})$. The p-values in Table 4.11 spans in the range of 0.016 to 0.079.

Regardless of the testing using simulated or real data, $S_1(HL)$, $S_1(MAD_n)$, $S_1(T_n)$, $S_1(S_n)$ and $S_1(Q_n)$ show the potential on preventing the inflation of Type I error. As in Table 4.8 using simulated data, all five proposed procedures are robust according to Guo and Luh's criterion of robustness and they generate the empirical Type I error rates below 0.05 level in general. Meanwhile, when using real data, the results on the proposed procedures (refer to Table 4.12) significantly show the difference in groups with their p-values below 0.05 level, together with the Kruskal-Wallis procedure. However, for simulated data, Kruskal-Wallis fails to meet Guo and Luh's criterion of robustness for negative pairing under four groups case. Furthermore, it is observed that when there is violation of normality, $S_1(MAD_n)$ and $S_1(T_n)$ are able to improve the performance of the original S_1 statistic in controlling Type I error rates for testing the equality of location parameters under unbalanced design for four groups case using both simulated and real data. Hence, from the results obtained, we can assume that the goal of the study has been achieved.

CHAPTER FIVE
CONCLUSION

5.1 Introduction

Statistical procedures are vital part in most industries nowadays such as in the field of education, medicine, business, economics, sciences and many more. There are many statistical procedures available in the form of software in the market today and they can be applied easily for analysis. However, the users must be aware of the assumptions attached to each procedure especially those under parametric statistics. Parametric statistics such as *t*-test and Analysis of Variance (ANOVA) commonly used as the statistical procedures for testing the equality of location measures for two groups and more than two groups cases respectively. Nonetheless, ANOVA and *t*-test are sensitive to certain assumptions for example normality and homoscedasticity. If these procedures violate the assumptions, Type I error will be affected which consequently will affect the result of the analysis. Hence, some researchers chose nonparametric statistics to overcome the issue. Indeed, nonparametric procedures are suitable to be applied without having to consider the distributional shapes but the procedures are not as powerful as parametric procedures. Therefore, another alternative that researchers can rely on is the robust procedures. This study focused on a robust procedure, S_1 statistic that was proposed by Babu *et al*. (1999) as alternative procedure when non-normality and/or variance heterogeneity occur. S_1 statistic uses median as the location parameter to measure the treatment effects across groups and it does not need to carry out any procedure of transforming or trimming the original data to attain symmetry. However, previous work on S_1 statistic failed to show robustness for most of the conditions under four groups case especially for extremely skewed distribution. Besides the issue of robustness, conservation Type I error rates (below 0.025 level) were generated by the proposed procedures in most conditions for balanced and unbalanced designs.

The main goal of this study is to improve the original S_1 statistic in testing the equality of location measures under skewed distributions. In order to accomplish the goal, S_1 statistic was modified by substituting the location parameter, median with Hodges-Lehmann estimator. A few robust scale estimators such as MAD_n, T_n, S_n and Q_n were selected to replace the default scale estimator of S_1 statistic. In addition, this study also looked into the possibility of improving the S_1 statistic by modifying the original scale estimator, $\hat{\omega}$; using the deviation of Hodges-Lehmann instead of median. Hence, the five proposed procedures are $S_1(HL)$, $S_1(MAD_n)$, $S_1(T_n)$, $S_1(S_n)$ and $S_1(Q_n)$.

In order to examine the effect of controlling Type I error when the assumptions of normality and variance homogeneity are violated, three conditions namely perfect, mild and extreme departures were created from the manipulation of several variables. These variables are number of groups (two and four groups), sample sizes (equal and unequal), groups variances (equal and unequal), distributional shapes (normal, chi-square, g-and-h = 0.5) and nature of pairings (positive and negative pairings). Normal distribution represented the zero skewness, while chi-square distribution with 3 degrees of freedom stood for moderate skewness and g = 0.5 and h = 0.5 distribution for extreme skewness. The proposed procedures were stimulated for 5000 times and then bootstrapped for 599 times for each condition. Type I error rates obtained from the proposed procedures were then compared with parametric, nonparametric and the original S_1 statistic. Before the best procedure can be determined, the robustness of each procedure was examined. According to Guo and Luh's criterion of robustness, a procedure is robust if its empirical Type I error rate falls below 0.075 level at $\alpha = 0.05$.

5.2 The S_1 Statistic

The five proposed procedures, $S_1(HL)$, $S_1(MAD_n)$, $S_1(T_n)$, $S_1(S_n)$ and $S_1(Q_n)$ were tested according to balanced (equal group sizes and equal group variances) and unbalanced designs (unequal group sizes and unequal group variances) under two and four groups cases. The purpose is to examine the effect of each procedure in controlling Type I error under different group designs and group cases for the condition of zero, mild and heavy skewness. Table 5.1 below presents the best and the worst procedures across all three levels of skewness for balanced designs under $J = 2$ and $J = 4$.

Table 5.1

The Best and the Worst Procedures for Balanced Design

Group Case	Distribution	Best Procedure	Worst Procedure
$J = 2$	Normal	$S_1(MAD_n)$ (0.0360)	$S_1(Q_n)$ (0.0266)
	Chi-Square	$S_1(S_n)$ (0.0272)	$S_1(Q_n)$ (0.0220)
	$g = 0.5$, $h = 0.5$	$S_1(MAD_n)$ (0.0212)	$S_1(HL)$ (0.0074)
$J = 4$	Normal	$S_1(S_n)$ (0.0234)	$S_1(Q_n)$ (0.0138)
	Chi-Square	$S_1(HL)$ (0.0128)	$S_1(Q_n)$ (0.0070)

	$g = 0.5$, $h = 0.5$		$S_1(MAD_n)$ (0.0096)	$S_1(HL)$ (0.0022)

In Table 5.1, we observe that $S_1(MAD_n)$ has the highest frequency which is three out of six conditions of being the best procedure among the five proposed procedures across groups cases especially for g-and-h = 0.5 distribution. Meanwhile, $S_1(S_n)$ emerges as the second best performer for balanced design. In contrast, $S_1(Q_n)$ is the worst procedure for normal and chi-square distributions whereas $S_1(HL)$ has the lowest Type I error rates for g-and-h = 0.5 distribution for balanced design regardless of the group case.

The best and the worst procedures for the condition of zero, mild and heavy skewness under unbalanced design for two and four group cases are summarized in Table 5.2 below.

Table 5.2

The Best and the Worst Procedures for Unbalanced Design

Group Case	**Distribution**	**Nature of Pairing**	**Best Procedure**	**Worst Procedure**
$J = 2$	Normal	Positive	$S_1(T_n)$ (0.0390)	$S_1(HL)$ (0.0288)
		Negative	$S_1(T_n)$ (0.0276)	$S_1(Q_n)$ (0.0140)
	Chi-Square	Positive	$S_1(T_n)$ (0.0460)	$S_1(HL)$ (0.0234)
		Negative	$S_1(T_n)$ (0.0402)	$S_1(HL)$ (0.0174)
	$g = 0.5$, $h = 0.5$	Positive	$S_1(T_n)$ (0.0304)	$S_1(HL)$ (0.0104)
		Negative	$S_1(MAD_n)$ (0.0200)	$S_1(HL)$ (0.0058)

$J = 4$	Normal	Positive	$S_1(MAD_n)$ (0.0252)	$S_1(Q_n)$ (0.0162)
		Negative	$S_1(MAD_n)$ (0.0196)	$S_1(Q_n)$ (0.0112)

Table 5.2 *continued*

	Chi-Square	Positive	$S_1(T_n)$ (0.0260)	$S_1(HL)$ (0.0148)
		Negative	$S_1(MAD_n)$ (0.0272)	$S_1(HL)$ (0.0158)
	$g = 0.5$, $h = 0.5$	Positive	$S_1(MAD_n)$ (0.0150)	$S_1(HL)$ (0.0028)
		Negative	$S_1(MAD_n)$ (0.0130)	$S_1(HL)$ (0.0032)

With regards to the shape of distribution and nature of pairings for unbalanced design, $S_1(T_n)$ is the best procedure for two groups case whereby $S_1(MAD_n)$ shows the best performance for four groups case. In the meantime, $S_1(HL)$ appears to be the worst procedure for both cases especially for chi-square and g-and-h = 0.5 distributions. In contrast, for balanced design, $S_1(HL)$ does not generate the lowest Type I error rate under chi-square distribution for $J = 2$ and $J = 4$. However, $S_1(Q_n)$ stays to be the worst for both group cases under normal distribution.

Overall, $S_1(MAD_n)$ emerges as the best procedure for balanced ($J = 2$ and $J = 4$) and unbalanced design ($J = 4$). For two groups case under unbalanced design, $S_1(T_n)$ has the best performance among five proposed procedures. In terms of Type I error rates, it can be clearly seen that the proposed procedures produce closest values to nominal level of 0.05 for unbalanced design compared to balanced design for two and four groups cases. Besides, the error rates of $J = 2$ are closer to the nominal level than the error rates of $J = 4$ for each group design. The five proposed procedures can be considered robust for every condition because all error rates fulfill Guo and Luh's criterion of robustness.

5.3 S_1 Statistic versus Parametric and Nonparametric Procedures

A comparison between the proposed procedures with parametric, nonparametric and the original S_1 procedures, $S_1(\hat{\omega})$ were also conducted in order to achieve the goal of this study. For parametric statistics, the chosen procedures were *t*-test and Analysis of Variance (ANOVA) representing the two and four group cases respectively. Meanwhile, for nonparametric procedures, Mann-Whitney was used for two groups case while Kruskal-Wallis was selected for the four groups case. Table 5.3 below summarizes the best procedure, the worst procedure and the range of Type I error rates for two groups, $J = 2$.

Table 5.3
Balanced and Unbalanced Designs for J=2

Group Design	Distribution	Nature of Pairing	Best Procedure	Worst Procedure	Range
Balanced	Normal		Mann-Whitney (0.0526)	$S_1(Q_n)$ (0.0266)	0.0266 to 0.0528
	Chi-Square		*t*-test (0.0500)	$S_1(Q_n)$ (0.0220)	0.0220 to 0.0566
	$g = 0.5$, $h = 0.5$		Mann-Whitney (0.0526)	$S_1(HL)$ (0.0074)	0.0074 to 0.0526
Unbalanced	Normal	Positive	Mann-Whitney and $S_1(\hat{\omega})$	$S_1(HL)$ (0.0288)	0.0288 to 0.0448

		(0.0448)		

Table 5.3 *continued*

	Negative	$S_1(\hat{\omega})$ (0.0422)	$S_1(Q_n)$ (0.0140)	0.0140 to 0.1268
Chi-Square	Positive	$S_1(T_n)$ (0.0460)	$S_1(HL)$ (0.0234)	0.0234 to 0.0666
	Negative	$S_1(T_n)$ (0.0402)	$S_1(HL)$ (0.0174)	0.0174 to 0.1678
$g = 0.5$, $h = 0.5$	Positive	Mann-Whitney (0.0426)	$S_1(HL)$ (0.0104)	0.0104 to 0.0426
	Negative	$S_1(MAD_n)$ (0.0200)	$S_1(HL)$ (0.0058)	0.0058 to 0.1048

As in Table 5.3, Mann-Whitney is the best procedure for $J = 2$ based on its highest frequency of producing the best Type I error rates among the nine conditions: three for balanced design and six for unbalanced design. Mann-Whitney shows the best performance particularly for normal and *g*-and-*h* = 0.5 distributions. However, its Type I error rates for negative pairings as shown in Chapter Four inflated to above 0.075 level for all levels of skewness. Similar situation happens to the parametric procedure, *t*-test for the same nature of pairing and distributional shapes. Inflation of Type I error rate will lead to spurious rejection of null hypothesis. In contrast, several S_1 procedures such as $S_1(\hat{\omega})$, $S_1(T_n)$ and $S_1(MAD_n)$ emerge as the better procedures for negative pairing under unbalanced design. Meanwhile, $S_1(HL)$ can be considered as the worst procedure for both group designs because the Type I error rates are at the furthest low from the nominal level of 0.05. The $S_1(Q_n)$ procedure has the second highest frequency of generating the lowest error rates among the nine conditions.

In terms of the range of Type I error rates, balanced design and positive pairings of unbalanced design have ranges that never exceed the value of 0.666. Thus, the intervals of values for negative pairings under unbalanced design go above 0.075 level across normal, chi-square and g-and-h = 0.5 distributions due to the inflated Type I error rates that are generated by t-test. Furthermore, the widest range for two groups case under negative pairings is 0.0174 to 0.1678 for chi-square distribution.

Table 5.4 below displays the similar pattern as Table 5.3 for four groups, $J = 4$.

Table 5.4
Balanced and Unbalanced Designs for J=4

Group Design	**Distribution**	**Nature of Pairing**	**Best Procedure**	**Worst Procedure**	**Range**
Balanced	Normal		Kruskal-Wallis (0.0498)	$S_1(Q_n)$ (0.0138)	0.0138 to 0.0510
	Chi-Square		ANOVA (0.0450)	$S_1(Q_n)$ (0.0070)	0.0070 to 0.0450
	$g = 0.5$, $h = 0.5$		Kruskal-Wallis (0.0498)	$S_1(HL)$ (0.0022)	0.0022 to 0.0498

Unbalanced	Normal	Positive	Kruskal-Wallis (0.0448)	$S_1(Q_n)$ (0.0162)	0.0162 to 0.0448

Table 5.4 *continued*

		Negative	$S_1(\hat{\omega})$ (0.0302)	$S_1(Q_n)$ (0.0112)	0.0112 to 0.2850
	Chi-Square	Positive	Kruskal-Wallis (0.0492)	$S_1(HL)$ (0.0148)	0.0148 to 0.0526
		Negative	$S_1(\hat{\omega})$ (0.0278)	$S_1(HL)$ (0.0158)	0.0158 to 0.2976
	g = 0.5, h = 0.5	Positive	Kruskal-Wallis (0.0498)	$S_1(HL)$ (0.0028)	0.0028 to 0.1492
		Negative	$S_1(MAD_n)$ (0.0130)	$S_1(HL)$ (0.0032)	0.0032 to 0.3554

According to Table 5.4, the nonparametric procedure, Kruskal-Wallis emerges as the best procedure as Mann-Whitney (J = 2) for the same conditions. It produces the nearest empirical Type I error rates to the nominal level of 0.05 for both group designs under J = 4. Again, parametric (ANOVA) and nonparametric procedures (Kruskal-Wallis) generate inflated Type I error rates that are above 0.075 level for negative pairings under unbalanced design for the condition of zero, mild and extreme skewness. In the meantime, for negative pairing under unbalanced design, two S_1 procedures, $S_1(\hat{\omega})$ and $S_1(MAD_n)$ show the best performance among all tested procedures. In contrast, $S_1(HL)$ performs the worst for both group designs under chi-square and g-and-h = 0.5 distributions. The $S_1(Q_n)$ procedure is the worst procedure under normal distribution regardless of group design.

Refer to the last column in Table 5.4, the inflated Type I error rates expand the ranges of $J = 4$ as compared to $J = 2$ across all three levels of skewness. The widest range is negative pairing of g-and-h = 0.5 distribution which spans from 0.0032 to 0.3554. Besides, inflated Type I error rate is found under positive pairing of g-and-h = 0.5 distribution, generated by ANOVA. This outcome is different from the parametric procedure for $J = 2$, as t-test does not cause the inflation of Type I error for the same condition.

Overall, we observe that nonparametric procedures show better control of Type I error rates compared to the five proposed procedures regardless of the group design and group case for normal, chi-square and g-and-h = 0.5 distributions. However, they have the tendency to produce inflated Type I error rates for negative pairings under unbalanced design for two groups and four groups cases. Parametric procedures also have the possibility of causing the inflation of Type I error as well due to the error rates that are larger than 0.075 level. In contrast, the five proposed procedures do not generate any inflated Type I error rates for all conditions. Furthermore, three proposed procedures, $S_1(MAD_n)$, $S_1(T_n)$ and $S_1(S_n)$ show an improvement on the original S_1 statistic in terms of controlling Type I error rates under skewed distributions for balanced and unbalanced designs under both cases. Among these three proposed procedures, $S_1(MAD_n)$ provides the best result (Type I error rates nearest to the nominal level of 0.05).

The main goal of this study is to improve the original S_1 statistic in testing the equality of location parameter for two groups and more than two groups cases when the distributions are skewed. In order to achieve the goal, five procedures based on S_1 statistic were proposed by integrating the statistic with some robust estimators. For the central tendency measure of S_1 statistic, the default median was substituted by Hodges-Lehmann. Besides, the original scale estimator of S_1 statistic was replaced by the variance of Hodges Lehmann and also four other robust scale estimators, namely MAD_n, T_n, S_n and Q_n.

Based on the analysis of results using stimulated data in Chapter Four, $S_1(MAD_n)$ emerges as the best performer among the five proposed procedures due to its closet Type I error rates to the nominal level of 0.05 with respect to the level of skewness. In order to examine the effectiveness of the procedure in controlling the Type I error, parametric, nonparametric and the original S_1 procedures were compared against the proposed procedures. For $J = 2$, *t*-test and Mann-Whitney were chosen while ANOVA and Kruskal-Wallis were selected for $J = 4$. Although the proposed procedures including $S_1(MAD_n)$ did not show better control of Type I error than parametric and nonparametric procedures for balanced and unbalanced designs under both group cases in general, these proposed procedures are robust for all conditions. In contrast, the parametric and nonparametric procedures are not robust for unbalanced design (negative pairing) under $J = 2$ and $J = 4$ across all three distributional shapes. Besides, the three proposed procedures, $S_1(MAD_n)$, $S_1(T_n)$ and $S_1(S_n)$ have shown some improvement over the original S_1 statistic by producing empirical Type I error rates that are closer to the nominal level under skewed distributions.

Other than using simulated data, the proposed procedures were also tested using real data and then compared against parametric (ANOVA), nonparametric (Kruskal-Wallis) and the original S_1 procedures. As explained in Chapter Four, students' exam scores were collected from four different lecturers of different classes teaching the same subject. These groups of data have equal group variances but different group sizes and distributional shapes. Three groups are normally distributed and one is not normal. Based on the significance results (p-value) obtained, Kruskal-Wallis performs the best followed closely by $S_1(HL)$ and $S_1(T_n)$. ANOVA is the only procedure that fails to show the significance of the test because its p-value is greater than the significance level of 0.05. Other than $S_1(HL)$ and $S_1(T_n)$, $S_1(MAD_n)$ also shows better performance than the original S_1 procedure, $S_1(\hat{\omega})$ when the comparison between the five proposed procedures and $S_1(\hat{\omega})$ is made.

The proposed procedures show robustness across all conditions as compared to the nonparametric Kruskal-Wallis and parmateric ANOVA. Even though Kruskal-Wallis shows the highest frequency in controlling Type I error among the investigated conditions, this procedure fails under certain conditions. Furthermore, three of the proposed procedures, $S_1(MAD_n)$, $S_1(T_n)$ and $S_1(S_n)$ show better performance than the original S_1 statistic in terms of simulated and real data study. Hence, the goal of the study has been achieved.

5.4 Suggestion for Future Research

It is hardly possible for researchers to collect real data which is always normally distributed as required by the parametric procedures for testing the equality of location parameter across groups. Therefore, the main concern of this study is to improve the original S_1 statistic, which is an alternative procedure for testing the equality of location parameters, when the assumptions of normality as well as variance homoscedasticity are not met. According to Guo and Luh's criterion of robustness, we have proved that the proposed procedures in this study are robust across all levels of skewness for balanced (equal group sizes and equal group variances) and unbalanced (unequal group sizes and unequal group variances) designs under two groups and more than two groups cases. Furthermore, the proposed procedures with robust scale estimators has improved the original S_1 statistic under extremely skewed distribution. However, the proposed procedures generate low Type I error rates especially for four groups case especially on $S_1(HL)$. Hence, there are two suggestions for future research to look into this problem. The first suggestion is with regards to the low Type I error on $S_1(HL)$, but the procedure produces good *p*-value on the real data result. Based on the overall results obtained using real data, we notice that $S_1(HL)$ produces better *p*-value than some proposed procedures such as $S_1(MAD_n)$ and $S_1(S_n)$. This outcome is in contrast to when the procedure was tested using simulated data, which $S_1(HL)$ generally performs the worst among all proposed procedures for every condition. Hence, we might be able to find the answer to this phenomena if further study on power is done. Secondly, using other robust estimators to replace the existing location and scale estimators S_1 statistic, the results on Type I error could be improved further. The study can also consider other number of group cases as well as different distributions for each group case.

REFERENCES

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

Bickel, P. J. (1965). On some robust estimates of location. *The Annals of Mathematical Statistics*, *36*(3), 847-858.

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

Boos, D. D. (1982). A test for asymmetry associated with the Hodges-Lehmann estimator. *Journal of the American Statistical Association*, *77*(379), 647-651.

Boos, D. D. and Monahan, J. F. (1986). Bootstrap methods using prior information. *Biometrika*, *73*(1), 77-83.

Box, G. E. P. (1953). Non-normality and tests on variance. *Biometrika*, *40*, 318-355.

Brownie, C and Boss, D. D. (1994). Type I error robustness of ANOVA and ANOVA on ranks when the number of treatments is large. *Biometrics*, *50*(2), 542-549.

Brunner, L. J. and Austin P. C. (2007). Inflation of Type I error in multiple regression when independent variables are measured with error. Retrieved from http://www.utstat.toronto.edu/~brunner/MeasurementError/MeasurementError3i.pdf

Chernick, M. R. (1999). *Bootstrap Methods: A Practitioner's Guide*. New Jersey, US: John Wiley & Sons, Inc.

Cui, H., He, X. and Ng, K. W. (2003). Asymptotic distributions of principal components based on robust dispersions. *Biometrika*, *90*(4), 953-966.

Dixon, W. J. and Yuen, K. K. (1974). Trimming and winsorization: a review. *Statistiche Hefte*, *2*, 157-170.

Donoho, D. L. and Huber, P. J. (1983). The notion of breakdown point. *In A Festscherift for Erich L. Lehmann* (Brickel, P. J., Doksum, K. & Hidges, J. L., Jr., eds), 157-184.

Fisher, R. A. (1935). *The Design of Experiments*. New York, US: Hafner.

Geyer, C. J. (2006). Breakdown point theory notes. Retrieved from http://www.stat.umn.edu/geyer/5601/notes/break.pdf

Gibbons, J. D. and Chakraborti, S. (2003). *Nonparametric Statistical Inference*, 4th ed. Florida, US: CRC Press.

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

Hall, P. and Sheather, S. (1988). On the distribution of a studentized quantile. *Journal of the Royal Statistical Society*, 381-391.

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

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J. and Stahel, W. A. (1986). *Robust statistics: The approach based on influence functions*. New York, US: Wiley.

Hodges, J. L. Jr. and Lehmann, E. L. (1963). Estimated of location based on rank tests. *The Annals of Mathematical Statistics*, *34*(2), 598-611.

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. (1964). Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, *35*(1), 73-101.

Huber, P. J. (1972). Robust statistics: a review. *The Annals Mathematical Statistics*, *43*(4), 1041-1067.

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

Kang, Y. and Harring, J. R. (2012). Investigating the impact of non-normality, effect size, and sample size on two-group comparison procedures: an empirical study. *Presented at the Annual Meeting of the American Educational Research Association (AERA),* SIG: Educational Statisticians, BC.

Kazempour, K. (1995). Impact of stratification imbalance on probability of Type I error. *The American Statistician*, *49*(2), 170-174.

Kellermann, A. P., Bellara, A. P., Gil, P. R. D., Nguyen, D., Kim, E. S., Chen, Y. and Kromrey, J. D. (2013). Variance heterogeneity and non-normality: how SAS PROC TEST can keep us honest. Retrieved from http://support.sas.com/resources/papers/proceedings13/228-2013.pdf

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

Keselman, H. J., Wilcox, R. R., Othman, A. R. and 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.

Khan, A. and Rayner, G. D. (2003). Robustness to non-normality of common tests for the many sample location problem. *Journal of Applied Mathematics and Decision Sciences*, *7*(4), 187.

Kruskal, W. H. and Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, *47*, 583-621.

Kulinskaya, E., Staudte, R. G., and Gao, H. (2003). Power approximations in testing for unequal means in a one-way ANOVA weighted for unequal variances. *Communications in Statistics – Theory and Methods*, *32*, 2353-2371.

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

Lix, L. M., Keselman, J. C. and 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 Education Research*, *66*(4), 579-619.

MacGillivray, H. L. and Cannon, W. H. (2002). *Generalizations of the g-and-h distributions and their uses*. Unpublished Thesis.

Maxwell, S. E. and Delaney, H. D. (2004). *Designing Experiments and Analyzing Data: A Model Comparison Perspective*, 2nd ed. Mahwah, NJ, US: Lawrence Erlbaumn Associates Publishers.

Md. Yusof, Z., Othman, A. R. and Syed Yahaya, S. S. (2010). Comparison of Type I error rates between T1 and Ft statistics for unequal population variance using variable trimming. *Malaysian Journal of Mathematical Sciences*, *4*(2), 195-207.

Morgenthaler, S. (1992). Least-absolute-deviations fits for generalized linear models. *Biometrika*, *79*, 747-754.

Nachar, N. (2008). The Mann-Whitney U: a test for assessing whether two independent samples come from the same distribution. *Tutorials in Quantitative Methods for Psychology*, *4*(1), 13-20.

Othman, A. R., Keselman, H. J., Padmanabhan, A. R., Wilcox, R. R. and Fradette, K. (2004). Comparing measures of the "typical" score across treatment groups. *British Journal of Mathematical and Statistical Psychology, 57*(2), 215-234.

Oshima, T. C. and 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.

Ott, R. L. and Longnecker, M. T. (2010). *An Introduction to Statistical Methods and Data Analysis.* Belmont, CA: Brooks/Cole Cengage Learning.

Pitman, E. J. G. (1937a). Significance tests which may be applied to samples from any population. *Royal Statistical Society Supplement*, *4*(1-2), 119-130.

Pitman, E. J. G. (1937b). Significance tests which may be applied to samples from any population. *Royal Statistical Society Supplement*, *4*(2), 225-232.

Pitman, E. J. G. (1938). Significance tests which may be applied to samples from any population: the analysis of variance test. *Biometrika*, *29*(3-4): 332-335.

Quenouille, M. H. (1949). Approximate tests of correlation in time-series. *Journal of the Royal Statistical Society*, *11*(1), 68-84.

Ramsey, J. B. (2001). *The Elements of Statistics: with Applications to Economics and the Social Sciences,* 1st ed. Boston, US:Cengage Learning.

Reed, J. F. (1998). Contributions to adaptive estimation. *Journal of Applied Statistics*, *25*(5), 651-669.

Ronchetti, E. M. (2006). The historical development of robust statistics. *In Proceedings of the 7th International Conference on Teaching Statistics (ICOTS-7).* Retrieved from https://iase-web.org/documents/papers/icots7/3B1_RONC.pdf

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

SAS Institute. (1989). *IML software: Usage and Reference, Version 6*, 1st ed. Cary, NC: SAS Institute.

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

Sheskin, D. J. (2011). *Handbook of Parametric and Nonparametric Statistical Procedures*, 5th ed. Florida, US: CRC Press.

Snedecor, G. W. and Cochran, W. G. (1980). *Statistical Methods*, 7th ed. U.S.A, Ames:Iowa State University Press.

Staudte, R. G. and Sheather, S. J. (1990). *Robust Estimation and Testing*. New York, US: Wiley.

Stigler, S. M. (1977). Do robust estimators work with real data? *The Annals of Statistics*, *5*(6), 1055-1098.

Stigler, S. M. (2010). The changing history of robustness. *The American Statistician, 64*(4), 277-281.

Sullivan, M. III (2004). *Statistics: Informed Decisions Using Data*. Upper Saddle River, NJ: Pearson Education, Inc.

Syed Yahaya, S. S. (2005). Robust statistical procedures for testing the equality of central tendency parameters under skewed distributions. Unpublished Thesis.

Syed Yahaya, S. S., Md Yusof, Z. and Abdullah, S. (2011). A robust alternative to ANOVA. *Proceedings of the Second International Soft Science Conference 2011 (ISSC 2011)*. Ho Chi Minh City, Vietnam.

Syed Yahaya, S. S., Othman, A. R. and Keselman, H. J. (2004). Testing the equality of location parameters for skewed distributions using S_1 with high breakdown robust scale estimators. *Theory and Applications of Recent Robust Methods, Series: Statistics for industry and technology*, 319 – 328.

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

Teh, S. Y., Md Yusof, Z., Yaacob, C. R. & Othman, A. R. (2010). Performance of the traditional pooled variance t-test against the bootstrap procedure of difference between sample means. *Malaysian Journal of Mathematical Sciences*, *4*(1), 85-94.

Tukey, J. W. (1960). A survey of sampling from contaminated distributions. *In Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling* (I. Olkin *et al*., eds), 448-485. Stanford University Press.

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

Wilcox, R. R. (1994). A one-way random effects model for trimmed means. *Psychometrika*, *59*(3), 289-306.

Wilcox, R. R. (1997). ANCOVA based on comparing a robust measure of location at empirically determined design points. *British Journal of Mathematical and Statistical Psychology*, *50*(1), 93-103.

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

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

Wilcox, R. R. (2012). *Introduction to Robust Estimation and Hypothesis Testing*, 3rd ed. San Diego, CA: Academic Press.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, *1*, 80-83.

Winter, J. C. F. D. and Dodou, D. (2010). Five-point likert items: t test versus mann-whitney-wilcoxon. Retrieved from http://pareonline.net/pdf/v15n11.pdf

Yi, G. Y. and He, W. (2009). Median regression models for longitudinal data with dropouts. *Biometrics*, *65*(2), 618-625.

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